MIDB: Multilingual Instruction Data Booster for Enhancing Cultural Equality in Multilingual Instruction Synthesis

Yilun Liu[1*], Chunguang Zhao[1*], Xinhua Yang[1*], Hongyong Zeng[1], Shimin Tao[1], Weibin Meng[1], Minggui He[1], Yan Yu[1], Hongxia Ma[1], Li Zhang[1], Daimeng Wei[1], Boxing Chen[2]

[1]Huawei, China
[2]Huawei Canada, Canada
{liuyilun3, zhaochunguang4, yangxinhua2}@huawei.com

Abstract

Despite doubts on data quality, instruction synthesis has been widely applied into instruction tuning (IT) of LLMs as an economic and rapid alternative. Recent endeavors focus on improving data quality for synthesized instruction pairs in English and have facilitated IT of English-centric LLMs. However, data quality issues in multilingual synthesized instruction pairs are even more severe, since the common synthesizing practice is to translate English synthesized data into other languages using machine translation (MT). Besides the known content errors in these English synthesized data, multilingual synthesized instruction data are further exposed to defects introduced by MT and face insufficient localization of the target languages, leading to cultural inequality in trained LLMs. In this paper, we propose MIDB, a Multilingual Instruction Data Booster to automatically address the quality issues in multilingual synthesized data. MIDB is trained on around 36.8k revision examples across 16 languages by human linguistic experts, thereby can boost the low-quality data by addressing content errors and MT defects, and improving localization in these synthesized data. Both automatic and human evaluation indicate that not only MIDB steadily improved instruction data quality in 16 languages, but also the instruction-following and cultural-understanding abilities of multilingual LLMs fine-tuned on MIDB-boosted data were significantly enhanced, suggesting an improved linguistic and cultural equality.

Code & Datasets — https://github.com/zhaocorey/MIDB

1 Introduction

Large language models (LLMs) have made significant strides in their performance in English, achieving impressive capabilities across a range of natural language processing (NLP) tasks (Achiam et al. 2023; DeepSeek-AI 2025). However, the multilingual abilities of most LLMs remain relatively underdeveloped (Lai, Mesgar, and Fraser 2024), particularly due to the predominance of English in the pretraining data used by many popular open-source LLMs, such as the LLaMA series (Touvron et al. 2023a). For example, in the case of LLaMA-2 (Touvron et al. 2023b), the ratio of non-English languages in its pretraining corpus is merely around 2%,

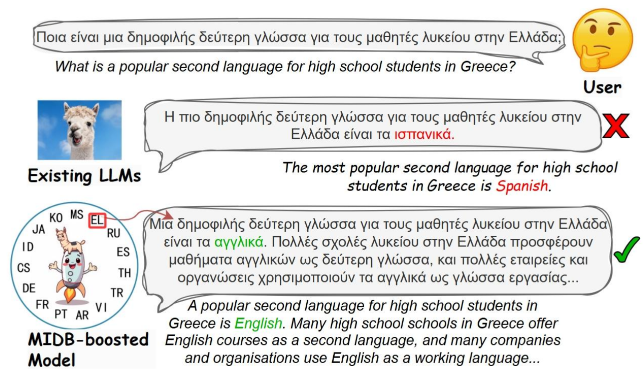


Figure 1: An example suggesting improved cultural equality in MIDB-boosted LLMs: it successfully identified a popular second-language in the cultural context of Greece. EL, RU, ES, *etc.*, are language codes. See code-name mapping of the 16 supported languages in Appendix D.1.[1]

which could significantly limit the abilities of downstream models to handle requests from diverse linguistic communities, and potentially leads to linguistic biases and insufficient cultural representations (Rystrøm, Kirk, and Hale 2025).

A widely-adopted mitigating approach is multilingual instruction tuning (IT), which introduces multilingual instruction pairs (*i.e.*, instruction-response pairs) during the post-training phase to enhance the model's ability to understand and process multiple languages (Chen et al. 2024b; Shaham et al. 2024). However, obtaining high-quality multilingual instruction data remains a significant challenge, especially for low-resource languages due to limited linguistic resources. Compared to monolingual contexts (*e.g.*, English), manually creating high-quality instruction datasets for multiple languages naturally incurs higher social utility costs and resource demands (Üstün et al. 2024). In contrast, using instruction synthesis techniques to automatically generate large-scale multilingual instruction data presents a more cost-effective and scalable alternative.

While research on instruction synthesis for English has flourished (Wang et al. 2023a; Taori et al. 2023; Xu et al.

[*]Equal contribution.

[1]Due to page limitation, all appendices are only available in arXiv version of this paper (Liu et al. 2025).

2024), multilingual instruction synthesis has largely relied on machine translation (MT) to adapt synthesized English datasets to target languages (Lai, Mesgar, and Fraser 2024; Chen et al. 2024b; Huo et al. 2025). A common approach involves translating the English Alpaca dataset (Taori et al. 2023), which contains 52k instruction pairs generated by LLMs, into other languages. Although MT provides a practical solution for rapid adaptation, it introduces data quality issues such as MT-introduced defects and insufficient cultural localization (as they tend to reflect cultures represented by English). As noted in prior studies (Zhou et al. 2024; Li et al. 2024; Liu et al. 2024), such data quality issues could significantly reduce the effectiveness of multilingual IT. Fully resolving these issues may require extensive human involvement, which is impractical for large scales—particularly in multilingual contexts.

To address this, we introduce MIDB, an automatic Multilingual Instruction Data Booster aimed at improving the quality of synthesized instruction data for multilingual IT. Inspired by data engineering techniques in English (Liu et al. 2024; Ge et al. 2024), where LLMs learn from human ratings or revision patterns, we collaborated with linguistic experts to build a dataset of 36.8k manual revision examples across 16 languages, including low-resource ones. These revisions target low-quality instruction pairs, correcting content issues (*e.g.*, accuracy, richness, relevance) and MT-induced defects (*e.g.*, fluency, correctness). These examples enable LLMs to learn human boosting strategies and enhance multilingual instruction data automatically. We also ensured that the dataset preserves language-specific features and emphasizes cultural and linguistic localization. This dataset is denoted as the Multilingual Expert Boosted (MEB) dataset.

We then used this MEB dataset to train MIDB to automate the enhancement of synthesized data quality for multilingual IT, leading to significant improvements on both multilingual data quality and model performance on multilingual and cultural-understanding abilities. In addition, since existing multilingual evaluation benchmarks for instruction-following abilities of LLMs also suffer from defects introduced by MT, as is observed by Chen et al. (2024c), we conducted manual localization of two most popular benchmarks, AlpacaEval (Dubois et al. 2023) and MT-Bench (Zheng et al. 2023), into 16 languages with the help of professional translators.

2 Social Impact of MIDB

MIDB addresses critical linguistic and cultural inequalities in multilingual AI systems, directly aligning with global priorities in social welfare and underserved communities. By tackling flaws in multilingual instruction synthesis, MIDB demonstrates potentials in two aspects:

(1) *Bridge the English-dominated Digital Divide.* As highlighted by a report by Stanford (Lynch 2025), The non-English speakers in the world may suffer a digital divide caused by English-dominated LLMs. “The ChatGPTs and Geminis of the world work well for the 1.52 billion people who speak English, but they underperform for the world’s 97 million Vietnamese speakers”, the report states. The impact for non-English communities may be beyond inconvenience, but a systematic exclusion from the gains by AI revolution, such as economic, educational and clinical opportunities, compared with those who are fluent in English. MIDB’s universal improvements on multilingual performance and released assets provide a cost-effective solution to this exclusion, enabling more localized AI applications (*e.g.*, LLM’s Performance in Vietnamese increased by 25.9% with MIDB).

(2) *Mitigating Cultural Inequality.* Current machine-translated instruction data (*e.g.*, from English) often erases cultural nuances, or even leads to cultural biases (*e.g.*, US-centric (Rystrøm, Kirk, and Hale 2025)). As shown in Fig. 1, Alpaca, an existing LLM, states that a popular second language in Greece is Spanish. This mistake may come from its English-originated training data, where Spanish is indeed a popular second language in the US. Such cultural inequalities may “collapsing cultural diversity into one big blob” (Lynch 2025), which is especially unsettling when LLMs are applied in high-stake scenarios like education. In contrast, MIDB’s focuses on cultural localization with human experts may help enhance cultural diversity in multilingual training data (*e.g.*, in Section 5.5, MIDB-boosted LLMs’ accuracy for cultural questions increased by 19.5% for five non-English cultures).

3 Related Work

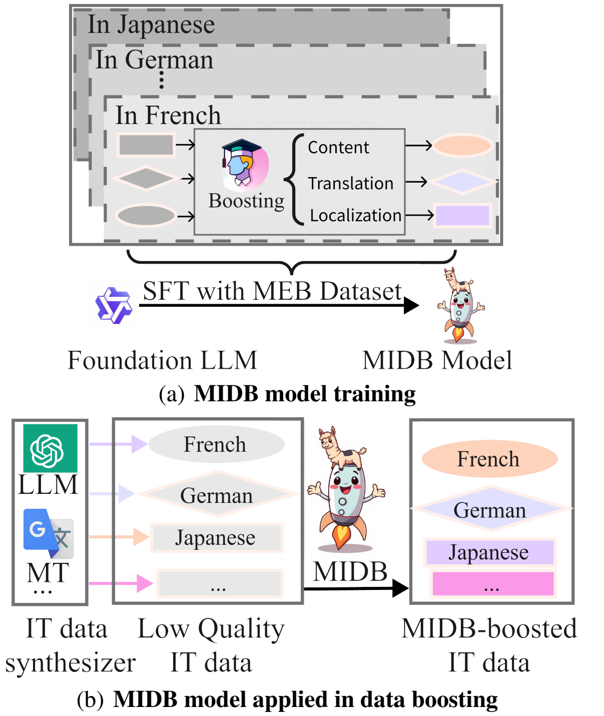
3.1 Multilingual Instruction Tuning

To improve the multilingual abilities of existing foundation LLMs, various training methods for enhancing multilingual IT are proposed. Chen et al. (2024b) combined multiple MT versions of the Alpaca dataset into training one multilingual model and achieved improved performance compared with monolingual baselines. Upon this combination, Zhu et al. (2023) further involved translation instructions into the training set, requesting the model to translate source sentences in English to sentences in target languages, to help transfer the knowledge learned in English. To make this transferring process more explicit, Zhang et al. (2024) composed specialized instruction pairs that ask the model to first process instructions in a pivot language (*e.g.*, English) and then produce response in the target language.

Despite improved training recipes, most existing methods directly adopt the synthesized Alpaca dataset and its MT versions as the training set, hampering effectiveness of their approaches due to known data quality issue. Our work fills this blank by focusing on data quality issue multilingual IT.

3.2 Instruction Data Synthesis

The technique of instruction data synthesis facilitates training of LLMs by generating instruction pairs using powerful LLMs, saving labors of human annotation. The pioneering attempt of instruction synthesis is Self-Instruct (Wang et al. 2023a), leveraging LLMs to produce instruction pairs given a small manually-written seed dataset. Subsequently, the Alpaca project (Taori et al. 2023) utilized the Self-Instruct strategy to generate 52k instruction pairs using LLMs. A lot of methods then focus on improving this pipeline, either by adding quality-filtering modules (Chen et al. 2024a; Ge et al. 2024) or increasing generation diversity (Xu et al. 2024; Liu et al. 2024). However, most existing technologies for instruction synthesis are targeted for the language of English,



(a) **MIDB model training**

(b) **MIDB model applied in data boosting**

Figure 2: Illustrations on (a) training stage and (b) inference stage of MIDB.

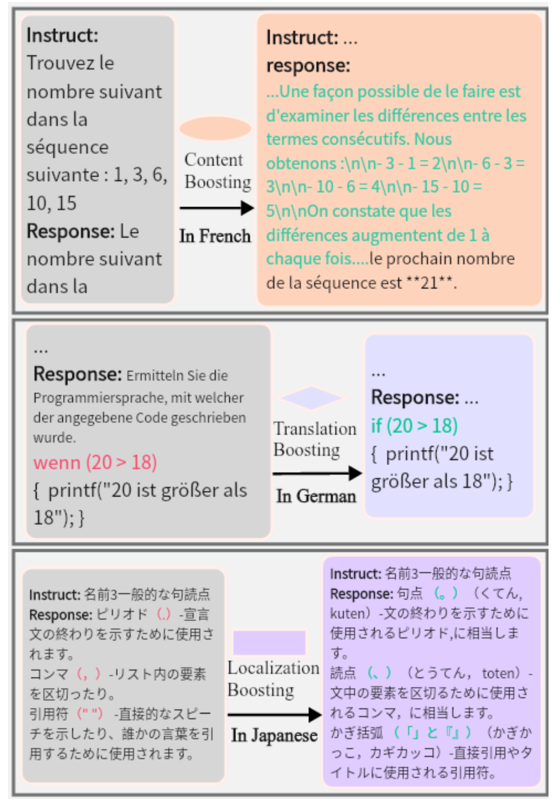


Figure 3: Typical issues addressed in MEB Dataset.

leaving the synthesis of multilingual instruction data underexplored. Compared with these existing methods, our work focuses on high-quality instruction synthesis for multiple languages.

We also noticed the concurrent work of Feng et al. (2025), who proposed a framework to synthesize cultural-related instruction pairs with critiques and translate them into multilingual data using MT. Our work, compared with theirs, works closely with language experts to solve the fine-grained problems such as fixing local expressions and cultural mismatching terms, potentially mitigating hallucinations by MT.

4 Methodology

The overview of the training and inference of MIDB is shown in Fig. 2. To construct training dataset for MIDB, a thorough manual correction and revision on subsets of the machine-translated Alpaca datasets (from English) was firstly conducted (Section 4.1-4.2). Section 4.3 discusses the training of MIDB. And Section 4.4 discusses the construction of our multilingual benchmarks.

4.1 Profile of Involved Multilingual Human Experts and Task Allocation Strategy

To accomplish the construction of the MEB dataset and two benchmarks, and human evaluation, we recruited and actively worked with a group of language experts (possessing an average of 6.5+ years' experience), coming from an international corporation's language service center with expertise in translation, localization, and editing. Experts were strategically allocated to the three tasks based on their proficiency while maintaining language coverage and avoiding bias through strict task separation. Quality was assured via a two-round review-rebuttal process and third-party oversight during annotation. Appendix C discusses full details of expertise, allocation strategies and quality control.

4.2 Building MEB Dataset

Preliminary Data Quality Study with Experts As the initial step, our language experts inspected several popular multilingual IT datasets, which were machine-translated from the synthesized English Alpaca Dataset (Taori et al. 2023). We then discussed with experts and identified several notable data quality issues within these samples and concluded them into three most common categories (Fig. 3 illustrates typical cases from each category, along with expert suggestions):

(1) *Content errors in the source* (e.g., *English) datasets.* Synthesized instruction datasets like Alpaca often contain content errors and defects due to LLM hallucinations during data generation (Chen et al. 2024a; Liu et al. 2024; Ge et al. 2024). These errors range from surface-level issues like formatting and grammatical mistakes (Ruebsamen 2023), to deeper defects such as logical inconsistencies, factual inaccuracies, and one-sided explanations (Liu et al. 2024). While some of these can be addressed with rule-based filtering, deeper issues still require human intervention. The first example in Fig. 3 illustrates a typical deeper defects, lack of comprehensiveness due to absence of intermediate process.

(2) *Defects introduced by MT.* Despite advances, current commercial MT systems or off-the-shelf LLMs still have significant defects in translation. For example, Lai, Mesgar, and Fraser (2024) found that multilingual Alpaca datasets translated via Google Translate API had an average error rate of around 30% in five low-resource languages. Such high translation error rates can lead to cascading issues. The second example in Fig. 3 addresses a common MT error, where the conditional statement "if" were mistakenly translated into the target language, potentially leading to catastrophic errors in the code compiler.

(3) *Insufficient Localization.* Direct translation of instruction data often leads to inadequate localization, as source instruction pairs tend to reflect the cultural and knowledge contexts of the source language (*e.g.*, English), which can lead to mismatches in the target language context. In the third case in Fig. 3, the response of a Japanese instruction, "List 3 common punctuation marks", mistakenly included English punctuation marks. With the help of language experts, the answer was accurately localized to incorporate Japanese-specific punctuation marks.

Criteria for Building MEB Dataset The categories derived from the typical issues above have been further summarized and listed as criteria in Table 1 for manual enhancement:

The "Content Boosting" category is primarily inspired by the criteria proposed by Liu et al. (2024), who introduced content revision criteria for English IT data and demonstrated their effectiveness through various experiments. Their English-based standards for content errors in synthesized data remain applicable in multilingual contexts. For example, criteria such as Relevance and Comprehensiveness are universally relevant and independent of the specific language. As such, we have inherited these criteria from their work.

The "Translation Boosting" category is derived from professional translation standards from the cooperated language service center, reflecting challenges for MT models such as "Omitted Translation" and "Translation Elegancy".

Localization-related criteria are the most challenging aspect of our work, mainly due to the limited availability of low-resource language experts and the lack of public instruction data. To address this, we propose four novel data-boosting criteria related to localization:

(1) *Cultural Relevance*: Adapting instruction pairs to reflect local culture, including references to local music, movies, and food where appropriate.

(2) *Geo-cultural Terms*: Recognizing that some entities are known by different names across regions. For instance, the Himalayas are referred to as Mount Everest in some languages. Instruction pairs should use terminology that aligns with these regional variations.

(3) *Ideological Localization*: This criterion addresses differences in religion, history, and local media, where the same input may yield different responses based on these factors. Some instruction pairs may need to be entirely restructured to align with these differences.

(4) *Local Expression*: Emphasizing the use of local, culturally relevant expressions instead of direct translations. These expressions help retain the unique features of each language, much like an ethnic costume showcases cultural identity.

Category	Ratio	Criteria	
Content Boosting	22.9%	Contextualization Feasibility Humanization Richness Readability	Relevance Timeliness Comprehensiveness Correctness Safety
Translation Boosting	24.4%	Fluency Translation Elegancy Spelling	Grammar Omitted translation Incorrect translation
Localization	52.7%	Culture localization Ideology localization	Geocultural term repair Expression localization

Table 1: Manual evaluation and enhancement criteria for quality issues in multilingual synthesized datasets.

Manual Enhancement Results A total of 16 languages were selected for inclusion in our study based on an evaluation of factors such as geographic coverage, popularity, impact, and available human resources. Notably, the selection includes four low-resource languages, ensuring coverage of all geographical regions (more details in Appendix D.2).

After dedicating over 485 person-days to the construction of the MEB dataset, we curated 36.8k manually boosted instruction pairs across the 16 languages, averaging approximately 2.3k pairs per language. As shown in Table 1, over 52.7% of the instruction pairs have been localized according to the proposed criteria, including both expression localization and cultural adaptation. And 22.9% and 24.4% of the manual enhancements address content issues and MT defects.

4.3 Training Design of MIDB

Building Training Samples for MIDB The manual enhancement examples are subsequently transformed into training samples for MIDB. As shown in Fig. 4, a training sample consists of *Prompt*, *Input* and *Output*. The *Prompt* is adapted from Liu et al. (2024), serving as a straightforward instruction for content refinement during the training of MIDB. The *Input* consists of an original instruction pair in the MEB Dataset, with its $< |\text{instruction}| >$, $< |\text{input}| >$, and $< |\text{response}| >$ concatenated into a string. The *Output* is the corresponding manually boosted instruction pair, serving as a learning target of MIDB to activate expert-aligned multilingual instruction boosting capabilities of the foundation LLM.

Joint Training Goal for 16 Languages To reduce deployment cost and facilitate connections between languages, we

Prompt	*Improve the following content to be more specific, detailed with more logical steps and grammarly corrected; avoid generating incorrect and misleading information in output; minimize hallucination in output.*
Input	*<|instruction|>XX1\n<|input|>XX2\n<|response|>XX3*
Output	*<|instruction|>YY1\n<|input|>YY2\n<|response|>YY3*

Figure 4: Template of MIDB's training samples.

trained MIDB as a unified model capable of boosting instruction pairs from all 16 languages. From the backbone θ, the joint training goal shown in Eq. (1) optimize it into θ_m, *i.e.*, MIDB. C_i represents the subset of the training samples from the i_{th} language within these 16 languages. For the j_{th} training sample in C_i, x_j is constructed by concatenating *Prompt* and *Input*, and y_j is directly the corresponding *Output*. Following Liu et al. (2024), only high-quality subsets are utilized in training MIDB with a quality-control coefficient α (defined as the top α-percent of samples that received the most revisions). See a sensitivity analysis on α and backbone θ in Section 5.7.

$$\theta_m = \arg\max_{\theta} \sum_{i \in [1,16]} \sum_{x_j \in C_i} \log P(y_j \mid x_j; \theta) \qquad (1)$$

4.4 Construction of Multilingual Test Sets

The evaluation encompasses three public benchmarks assessing instruction-following, multi-turn dialogue, and culture-specific understanding ability: AlpacaEval (Li et al. 2023), MT-Bench (Zheng et al. 2023), and BLEnD (Myung et al. 2024). AlpacaEval is a benchmark validated by 20k+ human judgments, aiming to assess the instruction-following ability of LLMs on general topics. MT-Bench includes 80 multi-turn questions across eight intent categories (*e.g.*, coding, reasoning, knowledge), designed to challenge strong models. AlpacaEval and MT-Bench were originally designed for English without multilingual support. We acquire the open-source English questions from AlpacaEval and MT-Bench, and a team of 20 professional experts (as discussed in Section 4.1) spent 175 person-days translating them to 16 languages. And hence we rename them to AlpacaEval-16L and MT-Bench-16L, respectively. Notably, the BLEnD benchmark already covers five languages within the scope of our evaluation.

5 Experiment

Implementation details of MIDB are introduced in Section 5.1. To ensure a comprehensive evaluation (setups discussed in Section 5.2), both data quality of MIDB-boosted datasets and performances of subsequently tuned LLMs are evaluated, encompassing both automatic (Section 5.3) and manual (Section 5.4) evaluations. Section 5.5 examines the enhancement on cultural understanding brought by MIDB. Section 5.6 verifies MIDB on an out-of-distribution dataset. In Section 5.7, we conduct an sensitivity analysis on critical parameters of MIDB.

5.1 Implementation Details

We explored different backbone models θ and different quality-control coefficient α for MIDB. In our main implementation of MIDB, we used LLaMA3.1-8B-Instruct (Grattafiori et al. 2024) as the backbone model, which has 8B parameters and possess strong multilingual abilities, and set quality-control coefficient to 30%. To efficiently adapt the backbone LLMs, we employed LoRA (Hu et al. 2021), a parameter efficient fine-tuning technique, with a rank of 64. MIDB was trained for three epochs with a learning rate of 4×10^{-4} and global batch size is set to 128. For training the instruction-following models, we utilized the same settings as the official Alpaca repository (Taori et al. 2023), with the exception of using different instruction datasets and using LLaMA3-8B as the backbone. The multilingual Alpaca datasets and Dolly datasets used during evaluation was translated by advanced LLMs from English. And a beam size of one was used for decoding across all models.

5.2 Evaluation Approach

LLM-as-judge Following recent studies (Ge et al. 2024; Liu et al. 2024; Chen et al. 2024a), we use advanced LLMs to automatically compare and score the responses of two candidates, based on the criteria from Zheng et al. (2023) (detailed prompt in Appendix E.1). This prompt asks the LLM to judge the helpfulness, relevance, accuracy, and detail of each response (which is better) and provide a rationale. However, this automatic method suffers from reported evaluation biases when the order of candidates is changed (Wang et al. 2023b). To alleviate the bias caused by positional deviation, we perform two evaluations swapping orders for each pair of candidates, and define the final judgment of a candidate as win only when it win twice, or win once and tie once (a win and a lose is counted as tie).

Human Seven experts independently rate each instruction pair or model response based on the criteria in Table 1, unaware of the sources of the evaluated samples. They assess the satisfaction level across predefined dimensions and give a comparison-based judgment on two candidate samples. However, human evaluation is inherently limited in terms of efficiency and scalability, due to its high cost and the need for domain expertise. Further more, due to limited resources of senior language experts, we are unable to conduct human evaluation for a larger set of languages. Therefore, our language selection for evaluation covers both high-usage and low-resource languages for coverage and representativeness.

Metrics Several metrics are utilized for presenting the judgments: (1) By default we use *win-lose-tie ratio*, which is given by $\frac{\#win}{\#all}$, $\frac{\#lose}{\#all}$ and $\frac{\#tie}{\#all}$, where $\#all$ is the number of test set samples; (2) For easy comparison between multiple baselines, we use *win rate*, formulated as $\frac{\#win+\#tie}{\#all}$; and (3) *winning score*, formulated as $\frac{\#win-\#lose}{\#all}+1$, to conveniently spot the winning side (score > 1).

5.3 Automatic Evaluation

Data Quality of MIDB-boosted Dataset We randomly sampled 520 data samples (strictly excluding those in MEB dataset) for each language from the machine-translated Alpaca dataset for quality assessment. An advanced LLM is employed to judge the winner for each original and MIDB-boosted data pair. Fig. 6 presents the win-lose-tie ratios of MIDB-boosted datasets in different languages. The results show that after MIDB boosting, all languages show a trend of significantly higher win ratios than lose ratios in terms of data quality. For example, Portuguese (PT) has a much higher winning ratio (46%) than its losing ratio (9%). This improvement suggests that the MIDB-boosted dataset primarily con-

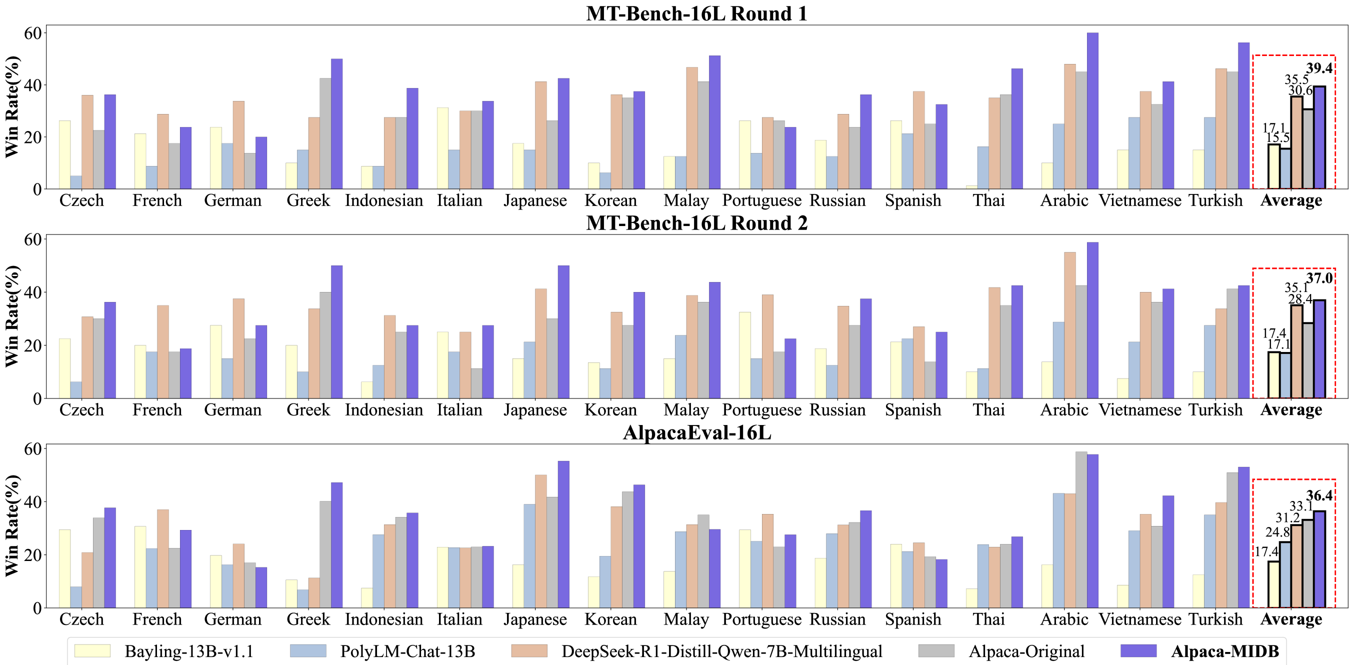


Figure 5: LLM-as-judge's evaluation on LLMs trained with MIDB-boosted/pre-boosted Alpaca datasets and 3 strong LLMs.

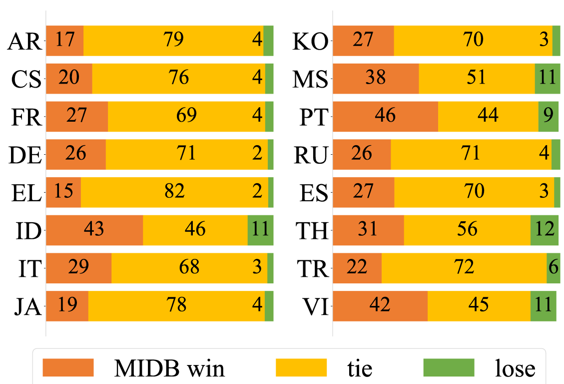


Figure 6: Win-lose-tie ratios of MIDB-boosted data compare with original data evaluated by LLM-as-judge.

tains high-quality instruction pairs, which can enhance LLM instruction tuning while maintaining the original dataset's integrity.

Evaluation of LLMs Tuned on MIDB-boosted Dataset Starting from the full machine-translated Alpaca datasets (with 52k samples) in 16 languages, MIDB boosted each sample and resulted in MIDB-boosted Alpaca datasets for each language. Then, we trained two groups of models: (1) Alpaca-Original models (Taori et al. 2023), trained on LLaMA-3-8B (Grattafiori et al. 2024) using the translated Alpaca datasets; (2) Alpaca-MIDB models, which are trained following the same setting as the Alpaca-Original models, but with the MIDB-boosted datasets replacing the translated Alpaca datasets. Several popular open-source LLMs that focus on multilingual tasks are also involved in the comparison. BayLing (Zhang et al. 2023) and PolyLM (Wei et al. 2023) are two multilingual LLMs with strong cross-lingual and multilingual performance on 10+ languages. DeepSeek-Multilingual (LightBlue 2025) is a recent model distilled from DeepSeek with extra training on multilingual data.

Following Section 4.4, we assess them on two datasets extended by our experts: the AlpacaEval-16L and MT-Bench-16L, covering evaluations on both multilingual instruction following and multi-turn dialogues. All models are compared with reference answers generated by LLaMA3.1-8B-Instruct, a strong multilingual baseline through extensive training.

As shown in Fig. 5, Alpaca-MIDB performs exceptionally well on both rounds of the MT-Bench-16L benchmark, achieving notably high scores in low-resource languages such as Thai and Greek. On AlpacaEval-16L, Alpaca-MIDB also maintains its leading position averagely. However, due to the varying proficiency of the foundation model (*i.e.*, LLaMA3-8B) across different languages, the performance of MIDB may exhibit fluctuations across certain language settings. Overall, Alpaca-MIDB exhibits strong multilingual instruction-following and multi-turn interactive capabilities, which originates from the high quality of its training data.

5.4 Human Evaluation

In addition to automatic evaluation, seven human experts (their profiles described in Appendix C) independently assessed the quality of data and the performance of subsequently tuned Alpaca-MIDB.

Data Quality We randomly selected 50 instruction pairs from the original Alpaca datasets with their MIDB-boosted

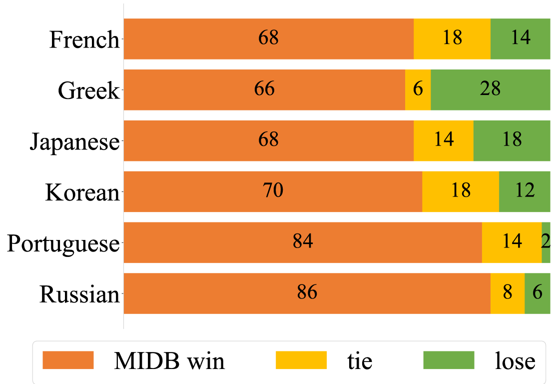


Figure 7: Win-lose-tie ratios of MIDB-boosted data compare with original data evaluated by human.

Language	AlpacaEval	MT-Bench	
		Round 1	Round 2
French	1.68	1.56	1.52
Greek	1.56	1.46	1.28
Japanese	1.72	1.68	1.40
Korean	1.36	1.38	1.38
Portuguese	1.62	1.84	1.88
Russian	1.68	1.72	1.52

Table 2: Winning scores by human on models trained with MIDB-boosted data *v.s.* original (score > 1 is win).

counterparts, and obtained independent ratings from multilingual experts who were unaware of the source of the samples. The results in Fig. 7 indicate that, after applying MIDB boosting, the data achieved a significantly higher average win ratios across all reviewers.

Alpaca-MIDB *v.s.* Alpaca-Original Human experts independently assessed the responses generated by Alpaca-MIDB and the original Alpaca-Original models on the AlpacaEval-16L and MT-Bench-16L test sets. The reviewers were unaware of the sources of the responses. As shown in Table 2, human reviewers consistently gave Alpaca-MIDB higher ratings (winning scores ranging from 1.28 to 1.72) compared with the original Alpaca-Original model. This strong multilingual performance of Alpaca-MIDB further confirms the effectiveness of the boosts made by MIDB, leading to improved user experience in multilingual contexts.

Analysis of Human Reviews It is worth noting that the enhancement brought by MIDB is consistently more pronounced across all 16 languages rated by human, as compared to ratings by LLM-as-judge. Thus, we conducted analysis on comments of reviewers and attributed it to nuanced but important improvements more perceivable to human, such as humanized tones and culturally appropriate expressions, which often leads to ties in LLM's judgments. Specifically, reviewers observed that responses of Alpaca-MIDB provided more detailed, human-like, well-structured, and readable content. Notably, Alpaca-MIDB was commented to have a richer reasoning process, particularly in addressing programming-related problems, resulting in superior outcomes quality.

Language	Original	MIDB-Boosted	Up ↑
Arabic	15.03	16.85	12.1%
Greek	18.72	22.03	17.7%
Spanish	25.00	28.49	14.0%
Indonesian	20.62	25.30	22.7%
Korean	18.50	24.19	30.8%

Table 3: Accuracy score (0-100) of Alpaca-Original and Alpaca-MIDB on cultural-specificity knowledge.

5.5 Evaluation of Alpaca-MIDB on Cultural Understanding Ability

English-centric LLMs often suffer from limited localization capabilities from other cultures, as the most of instruction pairs used during training reflect the contexts of English, leading to linguistic and cultural bias when applied in multilingual environments. To verify the effectiveness of MIDB on enhancing cultural understanding, we employed BLEnD (Myung et al. 2024), a manually constructed question-answering benchmark specifically designed to evaluate LLMs' understanding of daily knowledge across diverse cultures and languages. The accuracy is calculated based on average performance under two prompts: 1) Directly ask the LLM to provide answers; 2) Add a role setting for the LLM to answer as a native in target culture. We use an advanced LLM to verify if the LLM's response matches the manually labeled entity for each question (Appendix E.2), and calculate the percentage of correct answers as the score.

As shown in Table 3, models trained on datasets boosted with MIDB achieve performance improvements ranging from 12.1% to 30.8%. These results highlight the effectiveness of MIDB in enriching training data with culturally relevant and localized knowledge, thereby significantly enhancing the models' ability to understand and generate content that is contextually appropriate for non-English speakers.

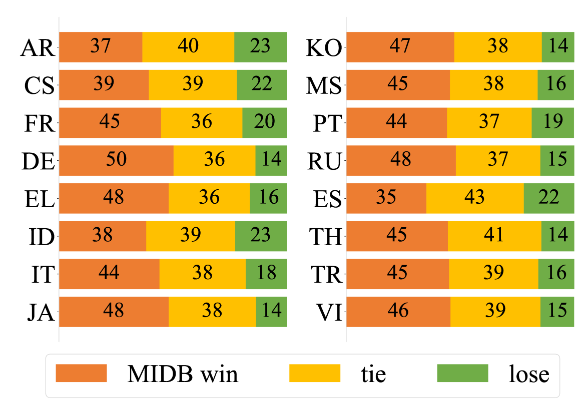
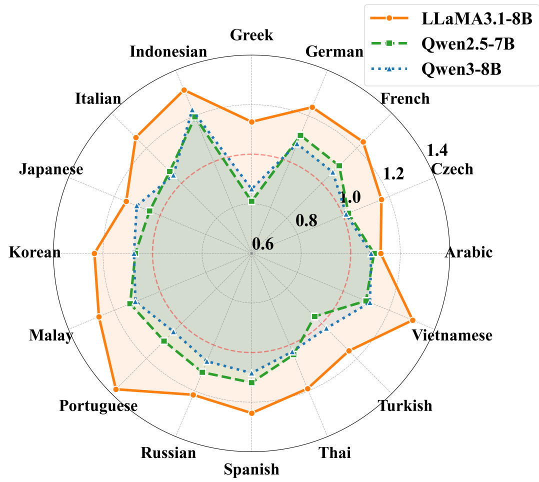
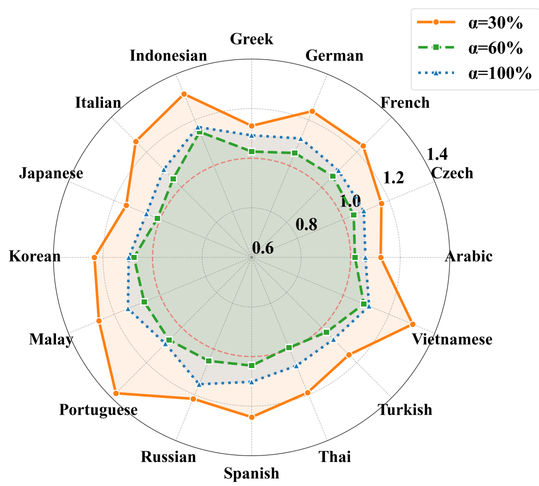


Figure 8: Win-lose-tie ratios of MIDB-boosted *v.s.* original data on Dolly-15k, an out-of-distribution dataset.



(a) Sensitivity on Varying Backbone Models θ



(b) Sensitivity on Quality-control Coefficient α

Figure 9: Performance of MIDB with (a) varying backbone models and (b) varying quality-control ratios, displayed as winning score of MIDB-boosted data against pre-boosted data randomly selected from Alpaca dataset. Score > 1 represents winning of MIDB (the dashed circle in red color).

5.6 Testing on Out-of-distribution Dataset

In addition to test MIDB on Alpaca-52k (in-distribution), we also select Dolly-15k (Conover et al. 2023), a dataset composed of real-world user instructions on complex tasks such as brainstorming and creative writing to evaluate the generalization capability of our method on out-of-distribution (OOD) data. Compared with Alpaca-52k, which is synthesized using LLMs, Dolly-15k possesses a different data distribution by collecting samples from human. In Fig. 8, despite OOD, the improvement of data quality after MIDB boosting remains significant, suggesting its strong generalization capabilities and promising application potentials.

5.7 Sensitivity Analysis

Variation on Backbone Model θ To investigate the impact of different backbone models θ on the performance of MIDB, we conduct a comparison training MIDB using different backbone models: LLaMA3.1-8B-Instruct and two additional models from Qwen families, *i.e.*, Qwen2.5-7B-Instruct (Yang et al. 2024) and Qwen3-8B-Instruct (Yang et al. 2025). The evaluation setting (*e.g.*, data sampling) is the same as Section 5.3. As shown in Fig. 9(a), the winning scores, when Qwen family models serve as the backbone, exceeds one in most languages, indicating a valid data enhancement brought by MIDB with varying backbones. In addition, due to its strong multilingual capabilities, LLaMA3.1-8B shows a significant advantages against Qwen models. Based on these findings, we select LLaMA3.1-8B-Instruct as the base model of MIDB for our main experiments.

Variation on Quality-control Coefficient α As discussed in Section 4.3, to ensure optimized performance of MIDB, we further incorporated the distance-based filtering mechanism in Liu et al. (2024) when curating training data for MIDB, which is to retain only high-quality samples with the largest edit distances between the expert-revised instruction pairs and the original counterparts. The coefficient α controls the percentage of this high-quality subset. Following Liu et al. (2024), we directly adopt α as 30%, which means that only the top 30% samples with the most revisions will be used as for training MIDB. Fig. 9(b) displays the performances of MIDB with varying α, using the evaluation setting in Section 5.3. Despite varying α, MIDB continuously achieves a winning score above one against original across languages, indicating a steady improvement on data quality after MIDB boosting. However, introducing more samples with less revision may lead to suboptimal performance of MIDB, as indicated by the relative advantages of "α=30%" group. This aligns with the hypothesis in (Liu et al. 2024) that manual revision samples with a higher edit distance contain more enriched learning patterns, while samples with a lower edit distance typically involve revisions limited to superficial aspects, such as grammar and layout adjustments.

6 Discussion

See discussions on limitations and ethical considerations in Appendix A and B.

7 Conclusion

In this paper, we proposed MIDB, an automatic tool to enhance linguistic and cultural equality in multilingual LLMs by addressing severe data quality issues such as cultural mismatching in synthesized instruction dataset. Experiment on 16 languages indicates universal improvement on data quality and subsequent model performance, suggesting a potential of MIDB to bridge the digital divide caused by English-centric AI technologies. Furthermore, the improvement on cultural understanding ability in the BLEnD test set renders MIDB a unique tool to mitigate cultural inequality concerns in high-stake areas such as education. Future work include expanding supported languages, testing on real-world scenarios and larger models.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chen, L.; Li, S.; Yan, J.; Wang, H.; Gunaratna, K.; Yadav, V.; Tang, Z.; Srinivasan, V.; Zhou, T.; Huang, H.; and Jin, H. 2024a. Alpagasus: Training a Better Alpaca Model with Fewer Data. In *International Conference on Learning Representations*.

Chen, P.; Ji, S.; Bogoychev, N.; Kutuzov, A.; Haddow, B.; and Heafield, K. 2024b. Monolingual or Multilingual Instruction Tuning: Which Makes a Better Alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, 1347–1356.

Chen, P.; Yu, S.; Guo, Z.; and Haddow, B. 2024c. Is It Good Data for Multilingual Instruction Tuning or Just Bad Multilingual Evaluation for Large Language Models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 9706–9726.

Conover, M.; Hayes, M.; Mathur, A.; Xie, J.; Wan, J.; Shah, S.; Ghodsi, A.; Wendell, P.; Zaharia, M.; and Xin, R. 2023. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. In *arXiv preprint arXiv:2501.12948*.

Dubois, Y.; Li, C. X.; Taori, R.; Zhang, T.; Gulrajani, I.; Ba, J.; Guestrin, C.; Liang, P. S.; and Hashimoto, T. B. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36: 30039–30069.

Feng, R.; Gao, S.; Chen, X.; Chen, L.; and Shang, S. 2025. CulFiT: A Fine-grained Cultural-aware LLM Training Paradigm via Multilingual Critique Data Synthesis. *arXiv preprint arXiv:2505.19484*.

Ge, Y.; Liu, Y.; Hu, C.; Meng, W.; Tao, S.; Zhao, X.; Xia, M.; Li, Z.; Chen, B.; Yang, H.; et al. 2024. Clustering and Ranking: Diversity-preserved Instruction Selection through Expert-aligned Quality Estimation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 464–478.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Huo, W.; Feng, X.; Huang, Y.; Fu, C.; Li, B.; Ye, Y.; Zhang, Z.; Tu, D.; Tang, D.; Lu, Y.; et al. 2025. Enhancing Non-English Capabilities of English-Centric Large Language Models through Deep Supervision Fine-Tuning. *arXiv preprint arXiv:2503.01275*.

Lai, W.; Mesgar, M.; and Fraser, A. 2024. LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback. *arXiv preprint arXiv:2406.01771*.

Li, M.; Zhang, Y.; Li, Z.; Chen, J.; Chen, L.; Cheng, N.; Wang, J.; Zhou, T.; and Xiao, J. 2024. From Quantity to Quality: Boosting LLM Performance with Self-Guided Data Selection for Instruction Tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 7595–7628.

Li, X.; Zhang, T.; Dubois, Y.; Taori, R.; Gulrajani, I.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. AlpacaEval: An Automatic Evaluator of Instruction-following Models. https://github.com/tatsu-lab/alpaca_eval.

LightBlue. 2025. DeepSeek-R1-Distill-Qwen-7B-Multilingual. https://huggingface.co/lightblue/DeepSeek-R1-Distill-Qwen-7B-Multilingual. Accessed: 2025-11-28.

Liu, Y.; Tao, S.; Zhao, X.; Zhu, M.; Ma, W.; Zhu, J.; Su, C.; Hou, Y.; Zhang, M.; Zhang, M.; et al. 2024. Coachlm: Automatic instruction revisions improve the data quality in llm instruction tuning. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, 5184–5197. IEEE.

Liu, Y.; Zhao, C.; Yang, X.; Zeng, H.; Tao, S.; Meng, W.; He, M.; Yu, Y.; Ma, H.; Zhang, L.; Wei, D.; and Chen, B. 2025. MIDB: Multilingual Instruction Data Booster for Enhancing Cultural Equality in Multilingual Instruction Synthesis. arXiv:2505.17671.

Lynch, S. 2025. How AI is leaving non-English speakers behind. *Stanford News*.

Myung, J.; Lee, N.; Zhou, Y.; Jin, J.; Putri, R.; Antypas, D.; Borkakoty, H.; Kim, E.; Perez-Almendros, C.; Ayele, A. A.; et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37: 78104–78146.

Ruebsamen, G. 2023. Cleaned Alpaca Dataset. https://github.com/gururise/AlpacaDataCleaned. GitHub repository.

Rystrøm, J.; Kirk, H. R.; and Hale, S. 2025. Multilingual!= multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in llms. *arXiv preprint arXiv:2502.16534*.

Shaham, U.; Herzig, J.; Aharoni, R.; Szpektor, I.; Tsarfaty, R.; and Eyal, M. 2024. Multilingual Instruction Tuning With Just a Pinch of Multilinguality. In *Findings of the Association for Computational Linguistics ACL 2024*, 2304–2317.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Üstün, A.; Aryabumi, V.; Yong, Z.; Ko, W.-Y.; D'souza, D.; Onilude, G.; Bhandari, N.; Singh, S.; Ooi, H.-L.; Kayid, A.;

et al. 2024. Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15894–15939.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2023a. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13484–13508. Toronto, Canada: Association for Computational Linguistics.

Wang, Y.; Yu, Z.; Zeng, Z.; Yang, L.; Wang, C.; Chen, H.; Jiang, C.; Xie, R.; Wang, J.; Xie, X.; et al. 2023b. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*.

Wei, X.; Wei, H.; Lin, H.; Li, T.; Zhang, P.; Ren, X.; Li, M.; Wan, Y.; Cao, Z.; Xie, B.; Hu, T.; Li, S.; Hui, B.; Yu, B.; Liu, D.; Yang, B.; Huang, F.; and Xie, J. 2023. PolyLM: An Open Source Polyglot Large Language Model. arXiv:2307.06018.

Xu, C.; Sun, Q.; Zheng, K.; Geng, X.; Zhao, P.; Feng, J.; Tao, C.; Lin, Q.; and Jiang, D. 2024. WizardLM: Empowering Large Pre-Trained Language Models to Follow Complex Instructions. In *International Conference on Learning Representations*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, Q. A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.-C.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; Qiu, Z.; Quan, S.; and Wang, Z. 2024. Qwen2.5 Technical Report. *ArXiv*, abs/2412.15115.

Zhang, S.; Fang, Q.; Zhang, Z.; Ma, Z.; Zhou, Y.; Huang, L.; Bu, M.; Gui, S.; Chen, Y.; Chen, X.; and Feng, Y. 2023. BayLing: Bridging Cross-lingual Alignment and Instruction Following through Interactive Translation for Large Language Models. arXiv:2306.10968.

Zhang, Z.; Lee, D.-H.; Fang, Y.; Yu, W.; Jia, M.; Jiang, M.; and Barbieri, F. 2024. PLUG: Leveraging Pivot Language in Cross-Lingual Instruction Tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7025–7046.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.

Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Zhu, W.; Lv, Y.; Dong, Q.; Yuan, F.; Xu, J.; Huang, S.; Kong, L.; Chen, J.; and Li, L. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.