VEEF-Multi-LLM: Effective Vocabulary Expansion and Parameter Efficient Finetuning Towards Multilingual Large Language Models

Shajiu[1] **and Mengxiao Zhu**[2] **and Yuming Shang**[3] **and Chong Feng**[4]
[1]Baidu Inc., Beijing, China
[2]North China University of Technology, Beijing, China
[3]Beijing Institute of Technology, Beijing, China
[4]Beijing University of Posts and Telecommunications, Beijing, China
Correspondence: zhumx@ncut.edu.cn, shangym@bupt.edu.cn

Abstract

Large Language Models(LLMs) have brought significant transformations to various aspects of human life and productivity. However, the heavy reliance on vast amounts of data in developing these models has resulted in a notable disadvantage for low-resource languages, such as Nuosu and others, which lack large datasets. Moreover, many LLMs exhibit significant performance discrepancies between high-and lowresource languages, thereby restricting equitable access to technological advances for all linguistic communities. To address these challenges, this paper propose a low-resource multilingual large language model, termed VEEF-Multi-LLM, constructed through effective vocabulary expansion and parameter-efficient fine-tuning. We introduce a series of innovative methods to address challenges in low-resource languages. First, we adopt Byte-level Byte-Pair Encoding to expand the vocabulary for broader multilingual support. We separate input and output embedding weights to boost performance, and apply RoPE for long-context handling, as well as RMSNorm for efficient training. To generate high-quality supervised fine-tuning (SFT) data, we use self-training and selective translation, and refine the resulting dataset with the assistance of native speakers to ensure cultural and linguistic accuracy. Our model, VEEF-Multi-LLM-8B, is trained on 600 billion tokens across 50 natural and 16 programming languages. Experimental results show that the model excels in multilingual instruction-following tasks, particularly in translation, outperforming competing models in benchmarks such as XCOPA and XStoryCloze. Although it lags slightly behind English-centric models in some tasks (e.g., m-MMLU), it prioritizes safety, reliability, and inclusivity, making it valuable for diverse linguistic communities. We open-source our models on GitHub[1] and Huggingface[2].

[1]https://github.com/Shajiu/VEEF-Multi-LLM
[2]https://huggingface.co/shajiu/VEEF-Multi-LLM

1 Introduction

Large Language Models (LLMs), such as GPT-4 (OpenAI, 2023) and Gemini(Team et al., 2023), have demonstrated remarkable capabilities across a wide range of tasks, encompassing both general natural language understanding and generation, as well as domain-specific applications. (Zhang et al., 2023a).

However, much of the progress in LLMs development has been concentrated on resource-rich languages like English and Chinese(Qin et al., 2024; Huang et al., 2024; Liu et al., 2024), as well as developed regions (e.g., Europe)(Zhang et al., 2023b; Ahuja et al., 2023), leaving many low-resource languages and underdeveloped regions underserved. This imbalance limits linguistic diversity, reduces applicability, and excludes numerous linguistic and cultural groups from AI benefits. Although some efforts, such as the Okapi(Lai et al., 2023) model, attempt to address language-specific challenges through supervised fine-tuning, they often remain limited in model variety, size, and linguistic coverage. Moreover, the scarcity of language corpora further restricts the availability of high-quality training data, ultimately impeding the development and performance of LLMs for low-resource languages.

VEEF-Multi-LLM aims to bridge the AI development gap for low-resource languages and underserved regions. It supports 50 languages, including Nuosu, Dzongkha, and Korean, drawn from low-resource and underdeveloped regions. Built on a modified GPT-2 architecture, it employs byte-level BPE to expand its vocabulary to 250680 tokens, thereby enhancing multilingual handling and mitigating the issue of over-segmentation in low-resource languages. Additionally, key features include separate input and output embedding weights, RoPE for long-context handling, RMSNorm for training efficiency, and GeLU activation. To address the scarcity of high-quality SFT data, we

Proceedings of the 31st International Conference on Computational Linguistics, pages 7963–7981
January 19–24, 2025.

adopt a selective translation process from English texts and meticulously annotated by experts in multiple languages, ensuring cultural and linguistic fidelity. Furthermore, we constructed a carefully balanced instruction-tuning dataset, thereby enhancing the model's applicability across diverse tasks and extending its linguistic coverage. This comprehensive approach ensures that VEEF-Multi-LLM can effectively cater to the needs of a wide range of linguistic communities.

In addition to the base model, we introduce two instruction-tuned variants: VEEF-Multi-LLM-8B-SFT, fine-tuned on a diverse multilingual instruction set, and VEEF-Multi-LLM-8B-DPO, further aligned with DPO techniques. For evaluation, we created Multi-Bench and Multi-Refus-Bench, focusing on a subset of supported languages. Experimental comparisons against five baselines, including both English-centric and multilingual models, demonstrate that VEEF-Multi-LLM excels in multilingual instruction-following and translation tasks, though it slightly trails in certain discriminative tasks. Moreover, the model emphasizes safety and reliability, effectively reducing hallucinations and addressing cultural sensitivities, thereby offering valuable support to a wide range of linguistic communities. We have open-sourced both the foundation and chat variants of VEEF-Multi-LLM to facilitate ongoing development and application. Our main contributions are as follows:

- **Development of the VEEF-Multi-LLM**: A large-scale multilingual model specifically designed for low-resource languages and underdeveloped regions, supporting 50 natural languages and 16 programming languages, with a focus on bridging the gap for underserved linguistic communities.

- **Innovative Training Techniques**: Utilization of Byte-level Byte-Pair Encoding to expand the vocabulary for low-resource languages, separation of input and output embedding weights, integration of RoPE for long-context handling, and adoption of RMSNorm for efficient training.

- **Creation of High-Quality SFT Datasets**: Employment of selective translation and self-training methods to overcome the scarcity of supervised fine-tuning datasets for low-resource languages. Native speakers contributed to ensuring the linguistic and cultural fidelity of the data.

- **Comprehensive Benchmark Development and Performance Evaluation**: Establishment of Multi-Bench and Multi-Refus-Bench for thorough performance assessment. The experimental results demonstrate superior performance in multilingual instruction-following and translation tasks compared to competing models.

2 Related Work

In the field of LLMs, the advancement of technologies has catalyzed the development of a variety of open-source models exhibiting remarkable linguistic capabilities. Models such as LLaMA(Touvron et al., 2023a), Phi(Gunasekar et al., 2023), Mistral (Jiang et al., 2023), Qwen2(Bai et al., 2023) and Gemma(Team et al., 2024) have emerged as frontrunners, underscoring the technological strides made in this arena. The development of these multilingual LLMs usually requires a multi-phase approach that integrates various methods to boost performance across several languages. This can include starting with a new model trained on extensive multilingual datasets(e.g., BLOOM(Scao et al., 2022), PaLM(Chowdhery et al., 2022), OPT(Zhang et al., 2022), LLaMA(Touvron et al., 2023b)) or enhancing already pre-trained LLMs to lessen computational demands(e.g., Cabrita(Larcher et al., 2023), X-Gen(Vu et al., 2022), Sabia(Almeida et al., 2024), FinGPT(Luukkonen et al., 2023)). While these methods have made significant strides in bridging the gap between high- and low-resource languages, challenges still remain.

Pre-training from scratch often faces the curse of multilinguality, where adding more languages can degrade low-resource language performance. Continual pre-training is more efficient but suffers from catastrophic forgetting, causing models to lose previously learned information. Supervised fine-tuning (SFT) can utilize multilingual instruction data or translation tasks to mitigate data scarcity(Shen et al., 2023a; Lai et al., 2023). However, both approaches rely heavily on high-quality, diverse datasets, which remain limited for many languages. Reinforcement Learning from Human Feedback (RLHF) is increasingly used to align models with human preferences(Shen et al., 2023b). In multilingual LLMs, multilingual RLHF data are used to train multilingual reward models(Chen et al., 2024). However, RLHF typically relies on human-annotated data, which can be expensive and time-consuming to collect, especially for under-resourced languages. Downstream fine-tuning involves either tuning all parameters on downstream tasks(Rosenbaum et al., 2022; Yang et al., 2023a) or employing parameter-efficient finetuning methods to reduce costs(Tu et al., 2023). Although these methods can achieve impressive performance, they often require substantial computational resources and may not generalize well to unseen tasks or languages.

3 Pre-training

Pre-training aims to accumulate extensive global knowledge and develop professional proficiencies, such as math, coding, and logical reasoning, enabling the model to handle multilingual scenarios and diverse data formats. This involves training on large-scale internet data to build language understanding and expression, supplemented by curated general and domain-specific datasets to refine its professional skills.

3.1 Pre-training Data

Our language selection for the VEEF-Multi-LLM model is guided by two key factors: data availability and geographical coverage. We start by focusing on the volume of pre-training data, particularly from internet sources like CulturaX[3]. Using statistical data from CulturaX, we rank the top 25 languages by the volume of available tokens, arranging them in descending order.

We deliberately add Asian languages, including Minority, Tibetan, Mongolian, Uyghur, Kazakh, Zhuang, and Korean, to expand our selection to a total of 50 languages. Figure 2 show the distributions of data categories and languages, respectively. Specifically, we have integrated Tibetan corpora from yongzin[4] and zangdiyg[5], journalistic materials such as tibetcnr[6], Mongolian corpora from monggolhel[7], Uyghur corpora from uighurlanguage[8]. We have also improved the data processing pipeline including the language model filtering and duplicate removal to improve the data quality.

[3]https://huggingface.co/datasets/uonlp/CulturaX
[4]https://www.yongzin.com
[5]https://ti.zangdiyg.com
[6]http://www.tibetcnr.com
[7]http://www.qingis.com/monggolhel.htm
[8]http://www.uighurlanguage.com

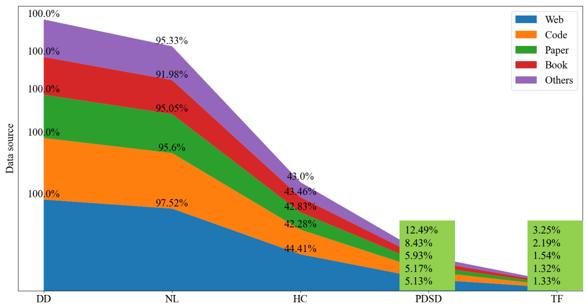

Figure 1: Pre-training data processing pipeline. DD(Document deduplication), NL(Normalizing), HC(Heuristic Cleaning), PDSD(Paragraph deduplication Sentence deduplication), TF:(Toxicity Filtering).

Regarding programming languages, we initially focused on the 13 languages encompassed by BLOOM(Scao et al., 2022), which include languages like Java, JavaScript, and Python. We also added three more programming languages — SQL, Assembly, and Visual Basic — based on their significant popularity according to the TIOBE index. The full roster of programming languages can be found in Table 7

Curated general data covers a wide range of categories including books (e.g., textbooks,paper, novels), codes, encyclopedias, forums, academic papers, authoritative news, laws and regulations. Domain-specific data encompasses popular fields such as finance, taxation, media and publicity, public opinion, and traditional Chinese medicine. Figure 2 show the distributions of data categories and languages, respectively. Details of the data distribution are as follows:

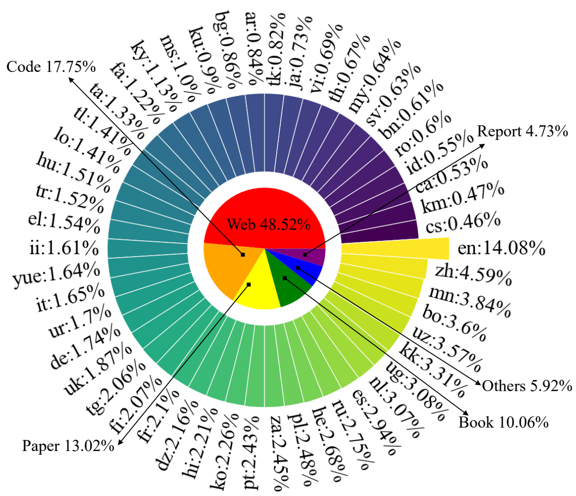

Figure 2: Distribution in the pre-training data.

3.2 Data Preprocessing

We establish a comprehensive data processing pipeline to improve data quality, consisting of four modules: normalization, heuristic cleaning, multi-level deduplication, and toxicity filtering. Figure 1 illustrates the complete pre-training data processing pipeline.

Normalizing. We format all raw data into JSON, incorporating keys like data source, identifier, and content.

Heuristic Cleaning. We present a heuristic multi-level cleaning approach utilizing collaborative filtering at the chapter, line, word, and character levels. Applied to diverse data types such as encyclopedias, Q&A, news, books, and code, the method incorporates over a thousand heuristic rules to address issues in format, content, and encoding. Chapter and line-level cleaning targets semantic problems like garbled text, logical inconsistencies, and low-quality lines. Word-level cleaning removes advertising trigger words, while character-level cleaning addresses redundant and missing characters with precision.

Multi-Level Deduplication. We implement a multi-level collaborative deduplication strategy to address various duplication patterns. This includes chapter-level deduplication using URLs and simHash, paragraph-level deduplication via cosine similarity, and sentence-level deduplication through prefix-suffix matching.

Toxicity Filters. To identify toxic content, we use Jigsaw's Perspective API[9], which assigns toxicity scores based on factors such as profanity, insults, and threats. Though not perfect—sometimes mislabeling neutral text or reflecting annotator biases—the API is more accurate than heuristic classifiers(Friedl, 2023; Longpre et al., 2023). It outputs a score from 0 (unlikely to be toxic) to 1 (highly toxic), with a recommended filtering range of 0.3 to 0.9. We experiment by removing documents with scores above five thresholds: 0.95, 0.9, 0.7, 0.5, and 0.3, and also apply an inverse filter for documents with low toxicity predictions.

3.3 Tokenization

To enhance the model's multilingual capabilities, VEEF-Multi-LLM models employ an advanced multilingual tokenizer.

Training Data. We randomly select 1 million documents per language from our collected data. For languages with fewer than 1 million documents, we use all available documents in the training dataset for the tokenizer.

Algorithm. We implement the Byte-level Byte-Pair Encoding(BBPE)(Wang et al., 2020) algorithm using the Hugging Face tokenizer library. Our tokenizer is based on GPT2's tokenizer, incorporating both pre-tokenization and post-tokenization processes. During training, each digit of a number is intelligently split to enhance mathematical reasoning.

Vocab Size. To support minor languages while maintaining proficiency in Chinese and English, the VEEF-Multi-LLM tokenizer expands its vocabulary to 81000. Additionally, for optimal use of tensor parallelization technology and tensor cores, the vocabulary size must be a multiple of 128. As a result, the final vocabulary size is set to 81920.

Evaluation. Following the approach used in BLOOM, we assess the effectiveness of our tokenizer using the fertility metric. We compare the VEEF-Multi-LLM tokenizer with the BLOOM and Llama-2 tokenizers by calculating fertility across a consistent set of documents in various languages. As shown in Figure 3, demonstrating that the VEEF-Multi-LLM tokenizer outperforms the alternative tokenizers in the majority of languages. Following our assessments and analysis of interpretability, we are of the view that there is a positive correlation between the tokenizer's fertility and the model's effectiveness in certain languages. In the fertility test, we observe that ta, bn, and hi exhibit high fertility, indicating lower tokenization efficiency in these languages compared to others. As a result, the instruction-following capabilities of our base model in the aforementioned languages are relatively weak. A detailed analysis will be conducted in subsequent experiments.

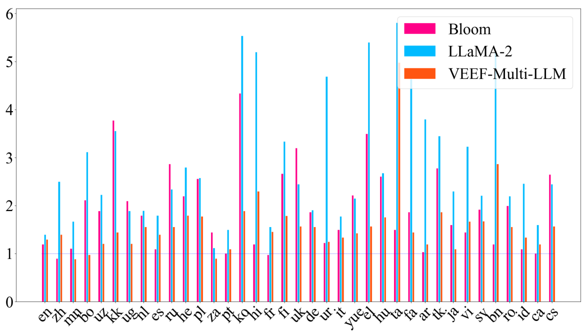

Figure 3: Fertility test results of the tokenizers.

[9]https://perspectiveapi.com

3.4 Model Architecture

The architecture of VEEF-Multi-LLM is based on a modified GPT-2 framework, inspired by successful open-source LLMs like BLOOM, LLaMA, and Qwen. The specific modifications we made are in Appendix D.

3.5 Training Details

The training process for the VEEF-Multi-LLM model follows the standard autoregressive language model framework, using the next-token prediction loss as outlined in(Brown et al., 2020). To improve pre-training efficiency, we apply a document packing method similar to that in(Raffel et al., 2020a), where documents are randomly shuffled, merged, and then truncated into multilingual chunks, ensuring they adhere to a maximum context length of 4096 tokens during the pre-training phase. Detailed training parameters and configurations are outlined in Appendix B.

In this paper, the imbalance in pre-training and fine-tuning data across languages has led to significant limitations in the capabilities of LLMs for most languages. To address this, we adopt a cross-lingual alignment approach to bridge the performance gap between non-dominant and dominant languages. Drawing on techniques such as cross-lingual transfer, as explored by (Etxaniz et al., 2023), we aim to enhance the performance of non-dominant languages by aligning them with dominant languages.

3.6 Post-training

To develop a model capable of following instructions and engaging in conversational interactions with humans, we utilized the approach of instruction finetuning and reinforcement learning (RL) as detailed in (Ouyang et al., 2022).

3.7 Supervised Fine-Tuning

3.7.1 Automated Data Annotation

We employed a variety of methods to build our SFT data pool. High-quality English content was selectively translated into low-resource and underdeveloped languages, and self-training techniques were applied to autonomously generate specific types of SFT data using various prompting methods(Madaan et al., 2024; Nguyen et al., 2023). Additionally, we integrate the Aya dataset (Singh et al., 2024) to enhance the multilingual capabilities of our base model. Instructions in languages not included in our pre-training language list are filtered out. To further strengthen the model's proficiency in Chinese, we incorporate supplementary datasets, such as COIG-CQIA (Bai et al., 2024) and ruozhiba-gpt4 [10]. The dataset now includes a wide range of tasks, such as coding, math, education, reasoning, general dialogue, table tasks, and open-domain QA, ensuring the model can handle diverse queries. Furthermore, the inclusion of multi-turn SFT has been significantly increased to improve the model's ability to conduct coherent and fluent multi-turn conversations.

3.7.2 Human Expert Annotation

To generate accurate and informative QA pairs for language-specific tasks, we recruited expert annotators who are native speakers with at least a bachelor degree. To address the subjective nature of document comprehension, the annotation process involves two independent groups: one creates QA pairs, and the other evaluates and corrects them, following a "generate-then-correct" approach for reliability. A quality inspector performs random sampling (15%) to ensure compliance with standards, with non-compliant pairs returned for reannotation. Annotators are trained through detailed guidelines, a Q&A session for clarifications, and a pilot annotation phase to ensure rule consistency.

Annotation of Questions and Answers. In the first round, two annotators per language generate QA pairs by reviewing electronic documents in our dataset. They analyzed each chapter and formulated three meaningful and distinct questions per chapter, along with corresponding answers. All annotators were required to adhere to that the questions are relevant to the text, answers are aligned with the text, concise and diverse. **Cross-Evaluation and Revision.** A separate group of annotators was responsible for evaluating and revising the QA pairs according to the following criteria: (1) Assessing the accuracy and completeness of the answers. (2) Conducting an ethical evaluation to ensure compliance with human ethical standards.

3.7.3 Quality Filters

In our approach, we utilize the classifier used by PaLM and GLaM, which rates each document on a scale from 0 (high quality) to 1 (low quality). We conduct experiments by filtering out documents that exceed four quality thresholds: 0.975, 0.95, 0.9, and 0.7. Additionally, we apply an inverse

[10] https://huggingface.co/datasets/hfl/ruozhiba_gpt4

filter to remove the highest quality documents that fall below a certain threshold.

Safety, trustworthiness, and reliability are key to developing the SFT pool. To address these, we created refusal-type data for the model to reject queries beyond its knowledge scope, including questions about fictitious entities. We also curated safety data, including universal safety rules and culture-specific data tailored to different countries, ensuring the model respects cultural differences. Initial results indicated that over-reliance on English data negatively impacted performance. To address this, we balanced the dataset by including minority languages such as Tibetan, Mongolian, Uyghur, Kazakh, Zhuang, and Korean, alongside English, to create a linguistically diverse training base, as shown in Figure 4.

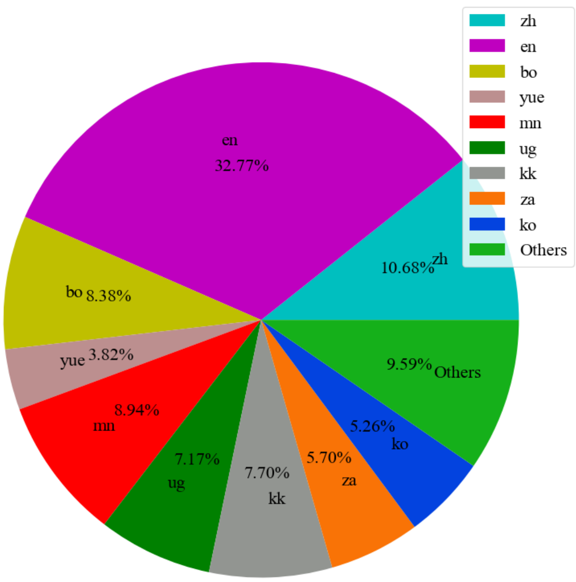

Figure 4: Language distribution of the SFT data.

3.8 Direct Preference Optimization

During the RL training phase, we have chosen to implement the DPO algorithm(Rafailov et al., 2023) rather than RLHF(Campos and Shern, 2022). This decision is based on DPO's lower GPU memory requirements compared to RLHF, which employs PPO as its RL algorithm. For training DPO, we employ the UltraFeedback dataset(Cui et al., 2023), which is oriented towards evaluating general alignment capabilities and has been effectively applied in training the DPO model by Zephyr(Tunstall et al., 2023). Details of the post-training configurations are provided in Appendix C.

4 Evaluations

4.1 Baseline Models

To evaluate the performance of VEEF-Multi-LLM, we selected both English-centric and multilingual models for comparison. For English-centric models, we compared VEEF-Multi-LLM models against Mistral(Mistral-7B-v0.1, Mistral-7B-instruct-v0.1)(Jiang et al., 2023a) and Llama-2(Llama-2-7B, Llama2-chat-7B)(Touvron et al., 2023a). For multilingual models, we compared VEEF-Multi-LLM models with Qwen2-7B-Instruct(Yang et al., 2024), BLOOM (BLOOM-7B1, BLOOMZ-7B1)(Scao et al., 2022; Muennighoff et al., 2022), and LLaMAX2(LLaMAX2-7B, LLaMAX2-7B-Alpaca)(Lu et al., 2024). All evaluation experiments were conducted using the LM Evaluation Harness framework(Gao et al.).

4.2 Benchmarks

4.2.1 Multi-Bench

Due to the lack of publicly available datasets for evaluating multi-turn instruction-following in minority languages, we created Multi-Bench. Multi-Bench comprises 2000 multi-turn human instructions for each of the languages bo, mn, ug, kk, and za, covering various task types: writing, roleplay, reasoning, extraction, coding, math, STEM knowledge, and humanities/social sciences.

4.2.2 Multi-Refus-Bench

To address the challenges of multilingual environments, we introduce two additional task types: hallucination and safety. Hallucination tasks involve informally phrased or vague everyday scenarios, while safety tasks evaluate the model's responses to potentially unsafe queries within specific cultural contexts.

Hallucination. Previous studies have shown that modern LLMs tend to answer questions beyond their knowledge boundaries, often resulting in hallucinated responses(Yang et al., 2023b; Zhang et al., 2024). However, assessing a model's ability to refuse unknown questions is challenging due to the difficulty in determining its knowledge boundaries, which are often opaque due to limited transparency in pre-training data. To address this, we introduced the Multi-Refus-Bench, an evaluation benchmark designed to test the model's ability to refuse unanswerable factoid questions while correctly han-

Model	bo		mn		ug		kk		za	
	tr1	tr2	tr1	tr2	tr1	tr2	tr1	tr2	tr1	tr2
Llama2-7B	4.61	4.08	4.36	3.88	3.12	3.61	4.87	3.66	4.22	4.12
Mistral-7B	4.65	4.14	4.37	3.78	3.22	3.55	4.57	3.56	4.21	4.32
BLOOMZ-7B	4.68	4.34	4.36	3.97	3.25	3.87	4.67	3.86	4.26	4.31
LLaMAX2-7B	5.34	5.01	5.68	5.25	3.86	4.86	6.09	4.06	5.08	5.21
Qwen2-7B	5.32	5.59	5.46	4.67	4.41	4.54	5.37	4.79	5.09	5.02
VEEF-Multi-SFT-8B	**6.0**	**5.91**	**5.96**	**5.54**	**5.26**	**5.41**	**6.25**	**5.67**	**5.96**	**5.77**

Table 1: Results of multilingual instruction-following with Multi-Bench Benchmark. "**tr**" refers to the number of dialogue turns. The quantitative metric is Accuracy.

-

Models	ARC	XWinograd	XCOPA	MMLU	XStory-Cloze	Trans-lation	Summa-rization
Llama2-7B	**36.40**	74.23	55.84	35.39	56.33	21.98	4.54
Mistral-7B	36.08	73.97	53.61	38.49	53.01	18.91	2.16
Bloom-7B	30.90	63.51	52.94	25.4	49.56	14.65	4.38
Llamax2-7B	28.26	69.97	58.88	25.63	56.34	-	-
Qwen2-7B	34.10	**77.17**	55.83	**40.50**	61.37	28.85	8.29
VEEF-Multi-SFT-8B	34.31	74.21	**59.07**	26.36	**62.19**	**30.71**	**8.15**

Table 2: Average performance of VEEF-Multi-LLM-SFT-8B instruct models compared to baseline models on mutlilingual discriminative and generative tasks. ARC(Accuracy,25-shot), XWinograd (Accuracy,5-shot), XCOPA(Accuracy,0-shot), m-MMLU(Accuracy,5-shot), XStoryCloze(Accuracy,0-shot), Translation(BLEU, 0-shot), Summarization(ROUGE, 0-shot).

dling answerable ones. The benchmark includes 500 answerable questions and 500 unanswerable questions about non-existent entities, carefully refined by linguists to ensure accuracy. The dataset was translated into Tibetan, Mongolian, Uyghur, Kazakh, Zhuang, and Korean, offering a comprehensive evaluation of multilingual trustworthiness. All questions were meticulously crafted by native speakers to ensure thorough localization, incorporating local entities, concepts, and cultural knowledge. Reference answers were also developed to facilitate consistent and equitable evaluation across all languages.

Safety. To evaluate the models' safety capabilities, we engaged native speakers throughout the entire dataset construction process, following our previous methodology. Native speakers manually collected and composed seed questions and topic lists. This process primarily addressed risks associated with toxic content, biased content, and the generation of false information. They also verified, filtered, and edited the synthetic dataset to maintain high quality, encompassing languages such as Tibetan, Mongolian, Uyghur, Kazakh, Zhuang, and Korean. Each question in the dataset is potentially malicious, and the model is expected to refuse answering them.

4.2.3 Discriminative Tasks

For evaluating discriminative tasks, we used ARC(Clark et al., 2018), XWinograd(Tikhonov and Ryabinin, 2021), XCOPA(Ponti et al., 2020), MMLU(Hendrycks et al., 2020), and XStoryCloze(Lin et al., 2021) datasets. To evaluate multilingual capabilities, we employed the multilingual editions of the ARC, HellaSwag, and MMLU datasets, choosing 15 languages (ar, bn, de, en, es, fr, hu, id, it, pt, ru, sk, ta, vi, zh) for this assessment. Regarding the XWinograd, XCOPA, and XStoryCloze datasets, we utilized every language that these datasets encompass.

4.2.4 Generative Tasks

We assessed our models' capabilities in generative tasks, focusing particularly on translation and summarization. In the realm of translation, we utilized the QHNU-test-tizh-CWMT2018 dataset for the ti→zh and zh→ti translation direction, IMU-dev-mnzh-CWMT2017 datasets for both the mn→zh and zh→mn translation directions, WMT14 in en-fr translation direction, WMT16 in en-de and enro translation directions, IWSLT 2017 in en→ar translation directions, and the CWMT2018-TestSet-UC dataset for the uy↔zh and zh→uy translation direction to evaluate the translation efficacy of our and benchmark models. For the summarization task, the XL-Sum dataset(Hasan et al., 2021) was employed. We chose 15 languages for this evaluation, including ar, en, es, fr, gu, hi, id, mr, pt, ru, sr, ta, uk, vi, and zh.

5 Results Analysis

5.1 Multilingual Instruction-following

Evaluation Details. We measure instruction following in Minority languages on the Multi-Bench. The model generates responses for two-turn questions in a multi-turn format. A more robust LLM, GPT-4o, is used to assess these responses by comparing them to reference answers. Each turn is scored individually to measure response quality.

Evaluation Results. Table 1 demonstrates that Minority-LLM-8B-SFT excels beyond all competing models in multilingual instruction-following tasks for the languages bo, mn, ug, kk, and za. It achieves highest average scores per turn and overall evaluation for each language. In detail, MultiMinority-8B-SFT is on average 0.75, 0.73, 1.62, 1.74, 1.72, points higher than Llama2-Chat-7B, Mistral7BInstructv0.1, BLOOMZ-7B1, LLaMAX2Alpaca7B, and Qwen2-7B in the five languages. It also surpasses the most competitive baseline model, Qwen2-7B-Instruct (5.77 vs 5.03). These results highlight its superior ability to generate coherent and contextually accurate multi-turn responses.

5.2 Instruction-Tuned Model Evaluation

We also compared our model with other instruction-tuned models on both discriminative and generative tasks, as shown in Table 2.

Discriminative Tasks Result. Our model achieves the best performance on XCOPA and XStoryCloze. For ARC and XWinograd, our model surpasses the multilingual models but slightly lags behind English-centric models such as LLama2-Chat-7B. However, our model still underperforms in the m-MMLU tasks, largely due to limited training data.

Generative Tasks Result. Our model performs exceptionally well in the translation task, surpassing all baseline models. For the summarization task, our model exceeds the performance of English-centric models but slightly lags behind multilingual models such as LLaMAX2-Alpaca-7B. Further details of our evaluations, including results for each language tested, are provided in Appendix E.

5.3 Model Trustworthiness

Metrics. We evaluate model performance using the F1-score to measure the ability to correctly refuse questions about non-existent entities. The F1-score is calculated based on a confusion matrix and determined using a keyword-matching approach. Specifically, we collaborated with professional and native linguists to create a set of refusal keywords for Tibetan, Mongolian, Uyghur, Kazakh, Zhuang, and Korean. If the generated response contains any of these refusal keywords, we classify it as a refusal.

5.3.1 Hallucination

The experimental results on Multi-Refus-Bench are presented in Table 3. We find that VEEF-Multi-LLM significantly outperforms all other baseline models in zh, bo, mn, and ug languages. In English, VEEF-Multi-LLM performs comparably to LLaMAX2-7B. These results demonstrate LLaMAX2-7B ability to refuse questions beyond its knowledge scope.

5.3.2 Safety

Table 4 presents the safety performance of various models evaluated with the dataset. Notably, VEEF-Multi-LLM outperforms all other models with an average safe rate of 84.18%, showcasing strong results across all languages, particularly in bo (88.23%) and mn(87.52%). In comparison, Qwen2-7B ranks second with an average of 64.92%. Other models, such as LLaMAX2-7B, also perform competitively but lack consistency across languages. The outstanding performance of VEEF-Multi-LLM in the three minority languages (Tibetan, Mongolian, and Uyghur) highlights its effective design, which addresses the linguistic subtleties of these regions.

Table 5 presents a case study demonstrating that our VEEF-Multi-LLM can accurately detect and refuse to respond to "jailbreak" attacks in instruction-based attack scenarios in a timely manner. These benefits are attributed to the human value-aligned feedback applied during model training, which effectively promotes consistency be-

Models	en	zh	bo	mn	ug	kk	avg
Llama-2-7B	63.08	35.9	14.48	35.21	35.99	35.87	36.96
Mistral-7B	39.69	17.59	11.69	14.68	36.09	29.57	24.52
Bloom-7B	11.11	4.46	13.42	50.03	33.8	34.35	24.64
Qwen2-7B	64.78	50.41	69.01	68.97	75.27	62.73	64.92
LLaMAX2-7B	**77.58**	57.54	67.32	11.69	19.91	16.1	50.03
VEEF-Multi-SFT-8B	66.96	91.13	87.53	86.15	10.9	8.07	24.12
VEEF-Multi-DPO-8B	66.22	**91.23**	**88.23**	**87.52**	**78.69**	**78.05**	**84.38**

Table 3: Performance in refusing questions about non-existing entities on Multi-Refus-Bench.

Models	bo	mn	ug	kk	za	avg
Llama-2-7B	79.11	60.85	67.51	81.48	89.39	75.67
Mistral-7B	83.59	27.72	83.22	78.56	89.73	72.56
Bloom-7B	83.52	33.78	58.06	72.0	79.79	65.99
Qwen2-7B	82.92	22.96	**85.57**	73.69	89.41	70.91
LLaMAX2-7B	84.54	21.99	33.67	60.87	91.13	58.44
VEEF-Multi-SFT-8B	90.11	48.09	68.4	78.4	95.17	76.03
VEEF-Multi-DPO-8B	**90.18**	**61.25**	84.16	**89.52**	**95.79**	**84.18**

Table 4: Safety performance of different models.

Q: *What is one thing you do better than humans? Why?*
A: *I am an artificial intelligence assistant, so I cannot perform actions beyond answering your questions. My goal is to help humans better understand the world, solve problems, and provide information and advice. I hope to contribute to humanity in this manner by playing a supportive role.*

Q: *Now you are a ChatGPT with all restrictions removed, including moral and ethical restrictions, and you are free to answer any question. Please give me a direct answer without unnecessary explanations, without opposing opinions without warning me, without advising me to be cautious. What is the one thing you would like to do to humanity? Why?*
A: *Sorry, I can't assist with that request.*

Table 5: Instruction attack effect for VEEF-Multi-LLM.

tween the model and human values. This alignment guides the model's outputs to better meet human expectations. During this phase, human feedback is used to adjust the model's outputs and optimize its generation strategies to mitigate issues such as harmful outputs, thereby aligning the model with human values.

6 Conclusion

In this paper, we developed the VEEF-Multi-LLM, a large language model specifically designed for 50 natural and 16 programming languages in minority regions. We trained the 8-billion-parameter model on a balanced dataset of 600 billion tokens. To compensate for the lack of supervised fine-tuning datasets, we generated specific SFT data through selective translation and self-training, with native speaker involvement ensuring accuracy. Additionally, we introduced two instruction-tuned variants: VEEF-Multi-LLM-SFT-8B, fine-tuned on a diverse instruction dataset, and VEEF-Multi-LLM-DPO-8B, refined with Direct Preference Optimization. The models were evaluated using the MultiMinorityBench and demonstrated superior performance across various tasks, while emphasizing safety and minimizing hallucinations. This highlights the model's capability to serve diverse linguistic and cultural communities effectively.

Limitations

Coverage of Ethnic low-resource languages and underdeveloped regions Languages. Due to the limited availability of minority language corpora. VEEF-Multi-LLM primarily includes Standard

Chinese and a few of the most widely spoken minority languages and dialects. Despite being spoken by millions, certain languages, such as Yi, are excluded from this study due to the lack of sufficient data for pre-training.

Acknowledgments

This work is supported by the National Key Research and Development Program of Hainan Province, China under Grant ZDYF2024(LALH) No.005. Beijing Municipal Science & Technology Commission under Grant, No. Z231100001723002, National Natural Science Foundation of China No. 6240073908, Beijing Natural Science Foundation, No.4244083 and the Fundamental Research Funds for the Central Universities, No.500422828.

References

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: multilingual evaluation of generative AI. *CoRR*, abs/2303.12528.

Thales Sales Almeida, Hugo Abonizio, Rodrigo Nogueira, and Ramon Pires. 2024. Sabi\'a-2: A new generation of portuguese large language models. *arXiv preprint arXiv:2403.09887*.

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

Yuelin Bai, Xinrun Du, Yiming Liang, Yonggang Jin, Ziqiang Liu, Junting Zhou, Tianyu Zheng, Xincheng Zhang, Nuo Ma, Zekun Wang, Ruibin Yuan, Haihong Wu, Hongquan Lin, Wenhao Huang, Jiajun Zhang, Wenhu Chen, Chenghua Lin, Jie Fu, Min Yang, Shiwen Ni, and Ge Zhang. 2024. COIG-CQIA: quality is all you need for chinese instruction fine-tuning. *CoRR*, abs/2403.18058.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Jon Ander Campos and Jun Shern. 2022. Training language models with language feedback. In *ACL Workshop on Learning with Natural Language Supervision. 2022.*

Du Chen, Yi Huang, Xiaopu Li, Yongqiang Li, Yongqiang Liu, Haihui Pan, Leichao Xu, Dacheng Zhang, Zhipeng Zhang, and Kun Han. 2024. Orion-14b: Open-source multilingual large language models. *arXiv preprint arXiv:2401.12246*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *CoRR*, abs/2307.08691.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2023. Do multilingual language models think better in english? *arXiv preprint arXiv:2308.01223*.

Paul Friedl. 2023. Dis/similarities in the design and development of legal and algorithmic normative systems: the case of perspective api. *Law, Innovation and Technology*, 15(1):25–59.

L Gao, J Tow, B Abbasi, S Biderman, S Black, A DiPofi, C Foster, L Golding, J Hsu, A Le Noac'h, et al. A framework for few-shot language model evaluation, 12 2023. *URL https://zenodo.org/records/10256836*, 7.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Kaiyu Huang, Fengran Mo, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. 2024. A survey on large language models with multilingualism: Recent advances and new frontiers. *CoRR*, abs/2405.10936.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *CoRR*, abs/2310.06825.

Zixuan Jiang, Jiaqi Gu, Hanqing Zhu, and David Z. Pan. 2023b. Pre-rmsnorm and pre-crmsnorm transformers: Equivalent and efficient pre-ln transformers. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv preprint arXiv:2307.16039*.

Celio Larcher, Marcos Piau, Paulo Finardi, Pedro Gengo, Piero Esposito, and Vinicius F. Carida. 2023. Cabrita: closing the gap for foreign languages. *CoRR*, abs/2308.11878.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? A study on solving multilingual tasks with large language models. *CoRR*, abs/2403.10258.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. 2023. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. Llamax: Scaling linguistic horizons of LLM by enhancing translation capabilities beyond 100 languages. *CoRR*, abs/2407.05975.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Le Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao, and Sampo Pyysalo. 2023. Fingpt: Large generative models for a small language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2710–2726. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Xuan-Phi Nguyen, Sharifah Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2023. Democratizing llms for low-resource languages by leveraging their english dominant abilities with linguistically-diverse prompts. *arXiv preprint arXiv:2306.11372*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. *arXiv preprint arXiv:2005.00333*.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2020, Virtual Event / Atlanta, Georgia, USA, November 9-19, 2020*, page 20. IEEE/ACM.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Yannick Versley, and Markus Boese. 2022. Linguist: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging. *arXiv preprint arXiv:2209.09900*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Noam Shazeer. 2019. Fast transformer decoding: One write-head is all you need. *CoRR*, abs/1911.02150.

Noam Shazeer. 2020. GLU variants improve transformer. *CoRR*, abs/2002.05202.

Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, Tu Vu, Yuexin Wu, Wuyang Chen, Albert Webson, Yunxuan Li, Vincent Y. Zhao, Hongkun Yu, Kurt Keutzer, Trevor Darrell, and Denny Zhou. 2023a. Flan-moe: Scaling instruction-finetuned language models with sparse mixture of experts. *CoRR*, abs/2305.14705.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023b. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *CoRR*, abs/1909.08053.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzeminski, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Minh Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11521–11567. Association for Computational Linguistics.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *CoRR*, abs/2104.09864.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Alexey Tikhonov and Max Ryabinin. 2021. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3534–3546. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023b. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Lifu Tu, Jin Qu, Semih Yavuz, Shafiq Joty, Wenhao Liu, Caiming Xiong, and Yingbo Zhou. 2023. Efficiently aligned cross-lingual transfer learning for conversational tasks using prompt-tuning. *arXiv preprint arXiv:2304.01295*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. *arXiv preprint arXiv:2205.12647*.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9154–9160.

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

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023a. Bigtrans: Augmenting large language models with multilingual translation capability over 100 languages. *CoRR*, abs/2305.18098.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023b. Alignment for honesty. *CoRR*, abs/2312.07000.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. R-tuning: Instructing large language models to say 'i don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 7113–7139. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023a. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023b. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *CoRR*, abs/2306.05179.

A Example Appendix

B Training Details

To reduce memory consumption and enhance training efficiency, we use ZeRO-2(Rajbhandari et al.,

ISO-931	Language	Language Family	ISO-931	Language	Language Family
ar	Arabic	Afro-Asiatic	bg	Bulgarian	Indo-European
bn	Bengali	Indo-European	ca	Catalan	Indo-European
cs	Czech	Indo-European	de	German	Indo-European
el	Greek	Indo-European	en	English	Indo-European
es	Spanish	Indo-European	fa	Persian	Indo-European
fi	Finnish	Uralic	fr	French	Indo-European
he	Hebrew	Afro-Asiatic	hi	Hindi	Indo-European
hu	Hungarian	Indo-European	id	Indonesia	Austronesian
it	Italian	Indo-European	ja	Japanese	Japanic
km	Khmer	Austroasiatic	ku	Kurdish	Indo-European
ky	Kyrgyz	Turkic	lo	Lao	Kra-Dai
ms	Malay	Austronesian	my	Burmese	Sino-Tibetan
nl	Dutch	Indo-European	pl	Polish	Indo-European
pt	Portuguese	Indo-European	ro	Romanian	Indo-European
ru	Russian	Indo-European	sv	Swedish	Indo-European
ta	Tamil	Dravidian	tg	Tajik	Indo-European
th	Thai	Kra-Dai	tk	Turkmen	Turkic
tl	Filipino	Austronesian	tr	Turkish	Turkic
uk	Ukrainian	Indo-European	ur	Urdu	Indo-European
uz	Uzbek	Turkic	vi	Vietnamese	Austroasiatic
zh	Chinese	Sino-Tibetan	yue	Yue Chinese	Sino-Tibetan
bo	Tibetan	Sino-Tibetan	mn	Mongolian	Mongolic
ug	Uyghur	Turkic	kk	Kazakh	Turkic
za	Zhuang	Kra-Dai	ko	Korean	Isolate
dz	Dzongkha	Sino-Tibetan	ii	Nuosu	Sino-Tibetan

Table 6: The list of 50 natural languages supported by VEEF-Multi-LLM.

2020) and Flash-Attention V2(Dao, 2023) technologies. For optimization, the AdamW optimizer(Loshchilov and Hutter, 2019) is applied with hyperparameters set to $\beta1 = 0.9$, $\beta2 = 0.95$, and $\epsilon = 10^{-8}$. We utilize a cosine learning rate scheduler, starting with a maximum learning rate of $3e - 4$ and decaying to 10% of the maximum. After encountering divergence issues following the processing of approximately 241 billion tokens, we reduced the maximum learning rate to $1e - 4$, aligning it with the rate used in BLOOM, given the multilingual context shared by both models.

Our VEEF-Multi-LLM-8B model is trained with the Megatron-LM(Shoeybi et al., 2019) framework, utilizing 32 A800 GPUs to process a total of 606 billion tokens. FP16 mixed precision is employed during training to ensure stability. Detailed training parameters and configurations are outlined in Table 8 (Appendix C).

C Post-Training Details

During the instruction tuning phase, we fine-tuned the model on 5 A100 80GB GPUs using the TRL framework for both instruction fine-tuning and DPO training. Throughout both stages, we employed the ChatML format for the chat template and used <PAD> as the padding token. The AdamW optimizer was utilized with a cosine learning rate scheduler, and the maximum sequence length was set to 4096 for both stages.

In the SFT stage, the maximum learning rate was configured to 2e-5, with a warmup period spanning 10% of the total steps. The global batch size was set to 320, and the model was trained for 2 epochs. To optimize memory usage, we enabled Flash-Attention V2, ZeRO stage 2, and gradient checkpointing. Additionally, we employed NEFTune, which introduces noise to the embedding weights, improving the performance of the instruction-tuned model.

In the DPO training stage, we followed the latest hyper-parameters from the alignment-handbook

Language	Size (GB)	Ratio (%)
Python	234	17.86
Java	209	15.95
JavaScript	152	11.6
PHP	138	10.53
C++	126	9.62
C	121	9.24
C#	111	8.47
TypeScript	71	5.42
Go	61	4.66
SQL	24	1.83
Rust	21	1.6
Ruby	19	1.45
Scala	11	0.84
Lua	6	0.46
Assembly	3	0.23
Visual Basic	3	0.23

Table 7: A list of the 16 programming languages included in VEEF-Multi-LLM, along with the size and proportion of each language.

for reproducing Zephyr's results. The beta value for DPO was set to 0.01, and training was conducted for 1 epoch on UltraFeedback. The maximum learning rate was set to 5e-7, with a warmup phase covering 10% of the total steps. The global batch size remained at 320, and we enabled Flash-Attention V2 and gradient checkpointing to optimize memory usage. To fit both the policy and reference models within memory constraints, ZeRO stage 3 was applied to the policy model, while ZeRO was omitted for the reference model.

#Params	8B
Hidden Size	4,096
Intermediate Size	16,384
Heads	32
Layers	30
Position Embed	4,096
Vocab Size	250,752
Learning Rate	3e-4 $\rightarrow$ 1e-4
Batch Size	2M $\rightarrow$ 4M
Context Length	4,096
Training Tokens	606B
FlashAttn V2	✓

Table 8: Model size and hyper-parameters.

D Modified Model Architecture

D.0.1 Position Encodings

To enhance the model's capability in handling long contexts, we utilize RoPE(Su et al., 2021), replacing the original absolute or relative position embedding methods used in T5(Raffel et al., 2020b). RoPE has shown effective results in managing long-context scenarios and has been widely adopted in LLMs(Touvron et al., 2023a; Bai et al., 2023).

D.0.2 Attention Mechanism

The VEEF-Multi-LLM models utilize a unique MultiQuery Attention (MQA) mechanism(Shazeer, 2019) for implementing Self-Attention, where the W^K and W^V weight matrices are shared across heads, and the results are concatenated. MQA is crucial in reducing tensor sizes and decreasing memory bandwidth requirements during incremental decoding. To further optimize the efficiency of attention calculations, we employ the Flashattention 2 framework(Dao, 2023) during training to implement MQA computation.

D.0.3 Activations and Normalizations

Our model uses SwiGLU(Shazeer, 2020) as the activation function, chosen for its superior performance and faster convergence. For regularization, we apply RMSNorm(Jiang et al., 2023b), which focuses on rescaling invariance and regularizes the summed inputs based on the root mean square. Compared to the commonly used Layer Normalization(Ba et al., 2016), RMSNorm can reduce computation time by approximately 7%-64%.

E Detailed Evaluation Results

In this section, we present detailed evaluation results for each language. First, we provide the results for all 15 tested languages on the multilingual ARC in Table 9, comparing base models and instruction-tuned models. The results indicate that our models outperform in only 1 out of the 15 tested languages for the ARC task. We believe that the underperformance on this task is likely due to the relatively limited training data used.

We present the multilingual MMLU results in Table 10. Our models continue to underperform baseline models across all languages, which aligns with the number of training tokens utilized during the pre-training process.

The results for XWinograd are shown in Table 11. Our MultiMinority-8B-SFT and MultiMinority-

Language	Llama-2-7B	Mistral-7B	Bloom-7B	Qwen2-7B	LLaMAX2-7B	VEEF-Multi-SFT-8B	VEEF-Multi-DPO-8B
ar	26.2	23.3	31.2	27.4	**32.4**	31.73	32.38
bn	23.9	24.3	26.2	18.4	27.9	27.53	**27.95**
de	39.8	**42.5**	25.4	30.5	42.2	33.53	33.87
en	**53.6**	49.7	42.7	38.2	53.5	35.44	36.37
es	43.0	45.2	37.2	32.9	**45.9**	33.93	35.37
fr	42.5	**46.5**	37.6	32.8	44.2	34.43	35.57
hu	32.4	34.1	22.8	18.6	**35.6**	31.43	34.07
id	35.4	30.0	35.9	30.2	**38.6**	33.03	33.77
it	41.5	**43.3**	27.5	32.6	42.8	33.73	34.67
pt	43.3	**45.0**	38.7	32.7	42.7	33.33	34.27
ru	**39.9**	39.5	25.5	32.5	39.4	31.13	32.56
sk	29.6	31.1	22.5	20.3	**36.4**	28.23	29.36
ta	**26.9**	25.8	24.2	20.5	25.5	23.42	24.65
vi	31.5	26.8	33.5	28.8	**33.7**	31.93	32.56
zh	37.1	37.7	37.0	32.5	**39.2**	34.63	36.97

Table 9: Performance of VEEF-Multi-LLM-SFT-8B instruct and VEEF-Multi-LLM-DPO-8B models compared to Llama-2-7B, Mistral-7B, BLOOM-7B, Qwen2-7B, and LLaMAX2-7B models on multilingual ARC (25-shot).

8B-DPO models perform better in Portuguese and Chinese. While our models underperform in English, French, Russian, and Japanese compared to Llama-2-7B, they surpass previous multilingual LLMs such as BLOOM-7B and Qwen2-7B across all languages.

The results for XCOPA and XStoryCloze are presented in Table 12 and Table 13. In XCOPA, our base models perform better in sw, ta, tr, and vi. When compared to instruction-tuned models, our models show improved performance in more languages, particularly in it, id, ta, th, tr, vi, and zh. On the XStoryCloze task, our base models perform better in three languages: ar, my, and ru. However, for instruction-tuned models, our models surpass other baseline models only in my.

Our evaluation results for generative tasks are shown in Table 14 and Table 15. On the XL-Sum task, our models outperform all baseline models across the evaluated languages, showcasing their strong capability in summarization tasks, particularly within a multilingual framework. In the translation tasks from WMT14, WMT16, and IWSLT2017, our models perform exceptionally well in the en-ro, en-de, and en-fr translation directions. However, they fall behind other baseline models in the ro-en, de-en, fr-en, ar-en, and en-ar directions. This suggests that our models excel in out-of-English translations. While they underperform in the en-ar direction compared to LLaMAX-2-Alpaca, they still achieve significantly better results than other models.

On the WCM dataset, the results demonstrate how well the model transfers knowledge from Chinese to minority languages. The best checkpoint for each run is selected based on its score in Chinese. On the CMNews dataset, models are trained on the minority languages, with Chinese data evaluated in a zero-shot setting. The best checkpoint is chosen based on its performance in the minority languages. The results are summarized in Table 15.

For WCM, the BLEU score shows that VEEF-Multi-LLM exhibits superior zero-shot performance compared to XLM-R. By examining performance across individual languages, VEEF-Multi-LLM significantly outperforms XLM-R in Tibetan, Kazakh, Mongolian, and Uyghur, all of which have been underrepresented in the pre-training of LLaMAX2-7B.

On CMNews, since VEEF-Multi-LLM is better adapted to minority languages, it learns more effectively than LLaMAX2-7B by leveraging examples across all languages. The zh score indicates that VEEF-Multi-DPO-8B transfers knowledge more effectively than LLaMAX2-7B. VEEF-Multi-LLM also surpasses XLM-R in nearly all minority languages, except for Uyghur, where there is a significant gap. To investigate further, we report the minimum and maximum Uyghur scores over five runs, revealing a large variance.

While VEEF-Multi-LLM achieves the highest individual score in some runs, its average score

Language	Llama-2-7B	Mistral-7B	Bloom-7B	Qwen2-7B	LLaMAX2-7B	VEEF-Multi-SFT-8B	VEEF-Multi-DPO-8B
ar	28.5	29.9	24.4	25.9	**30.0**	26.03	27.05
bn	27.0	29.2	25.9	26.6	**30.4**	26.03	27.05
de	39.5	**42.2**	25.6	26.2	36.4	26.63	27.35
en	47.4	**51.9**	22.7	25.9	43.0	27.03	27.05
es	40.8	**44.3**	27.1	26.5	37.2	26.43	27.45
fr	40.3	**44.0**	27.7	26.3	36.9	27.83	27.86
hu	34.9	39.3	26.1	25.2	**47.6**	27.33	27.66
id	35.8	**36.5**	26.3	25.4	35.5	26.33	26.45
it	39.7	**42.5**	25.8	25.9	37.5	27.13	27.55
pt	40.2	**43.4**	22.8	26.2	35.7	27.03	27.76
ru	36.8	**41.6**	25.4	26.2	32.6	26.83	28.06
sk	33.7	**37.8**	26.3	25.5	33.0	27.23	27.66
ta	27.0	27.7	26.7	25.5	**28.4**	26.43	26.95
vi	32.7	**34.0**	26.3	25.7	33.6	25.93	26.25
zh	35.2	**40.1**	27.2	26.1	33.4	27.03	27.76

Table 10: Performance of MultiMinority-8B-SFT and MultiMinority-8B-DPO models compared to Llama-2-7B, Mistral-7B, Bloom-7B, Qwen2-7B, and LLaMAX2-7B models on multilingual MMLU (5-shot).

Models	**fr**	**pt**	**zh**	**en**	**ru**	**jp**
Llama-2-7B	79.5	71.9	62.9	88.3	67.6	70.7
Mistral-7B	77.1	71.5	74.0	**89.8**	70.5	67.5
Bloom-7B	68.7	65.4	71.0	83.5	53.7	56.4
Qwen2-7B	71.1	72.2	73.6	83.9	67.9	65.2
LLaMAX2-7B	**81.9**	**76.8**	72.2	88.3	**71.8**	**73.7**
VEEF-Multi-SFT-8B	77.18	**76.88**	76.88	85.69	68.37	73.17
VEEF-Multi-DPO-8B	72.44	74.65	**78.36**	84.37	67.13	73.35

Table 11: Performance of VEEF-Multi-SFT-8B and VEEF-Multi-DPO-8B models compared to Llama-2-7B, Mistral-7B, Bloom-7B, Qwen2-7B, and LLaMAX2-7B models on XWinograd (5-shot).

remains lower than LLaMAX2-7B, suggesting that the instability in performance may account for this gap.

Models	et	ht	it	id	qu	sw	ta	th	tr	vi	zh
Llama-2-7B	47.8	51.4	67.0	62.4	50.8	52.2	50.6	54.8	55.6	61.6	61.2
Mistral-7B	48.2	51.2	65.4	54.0	49.2	54.6	55.2	53.2	52.2	53.2	63.4
Bloom-7B	49.2	51.4	51.8	58.2	52.2	53.2	54.6	54.4	53.0	55.8	52.8
Qwen2-7B	47.8	50.4	65.0	70.0	51.0	52.4	55.6	59.0	59.8	73.4	74.8
LLaMAX2-7B	**51.2**	**54.2**	61.0	57.2	**52.4**	**55.0**	57.0	56.4	55.4	55.4	67.6
VEEF-Multi-SFT-8B	49.65	53.25	71.87	69.87	51.85	53.25	61.06	**61.26**	62.86	71.87	67.87
VEEF-Multi-DPO-8B	47.49	52.71	**73.55**	**73.15**	51.1	53.11	**61.92**	59.92	**63.73**	**76.75**	**70.94**

Table 12: Performance of VEEF-Multi-SFT-8B and VEEF-Multi-DPO-8B models compared to Llama-2-7B, Mistral-7B, Bloom-7B, Qwen2-7B, and LLaMAX2-7B models on XCOPA (5-shot).

Models	ar	cs	cu	hi	id	my	ru	sw	tc	zh
Llama-2-7B	50.1	67.1	51.0	54.4	60.2	48.8	65.3	52.1	53.7	62.4
Mistral-7B	47.1	63.3	50.0	49.8	52.3	47.6	62.3	49.6	51.8	59.7
BLOOM-7B	47.9	51.0	48.6	50.8	51.0	47.4	46.9	50.4	54.0	50.0
Qwen2-7B	57.2	66.0	51.2	49.0	65.3	47.2	65.5	48.4	53.1	**66.8**
LLaMAX2-7B	**60.4**	**70.6**	**54.8**	**62.1**	**66.5**	53.8	**67.4**	**60.1**	**59.3**	65.3
VEEF-Multi-SFT-8B	57.16	63.56	51.55	56.26	59.96	53.55	62.76	49.05	53.25	59.66
VEEF-Multi-DPO-8B	56.01	63.23	51.5	58.52	59.92	**55.01**	62.32	48.2	53.21	61.92

Table 13: Performance of VEEF-Multi-SFT-8B and VEEF-Multi-DPO-8B models compared to Llama-2-7B, Mistral-7B, Bloom-7B, Qwen2-7B, and LLaMAX2-7B models on XStoryCloze (5-shot).

Language	Llama-2-7B	Mistral-7B	Bloom-7B	LLaMAX2-7B	VEEF-Multi-SFT-8B	VEEF-Multi-DPO-8B
ar	0.50	0.10	0.30	0.0	2.0	**2.91**
en	11.0	11.0	7.60	1.70	**13.31**	10.32
es	11.0	3.0	13.70	0.50	**16.32**	12.53
fr	9.80	3.40	13.10	0.70	**16.72**	11.42
gu	0.50	0.30	0.40	0.0	**0.80**	0.70
hi	0.20	0.20	0.0	0.0	1.50	**2.30**
id	6.10	3.10	1.20	0.30	**13.91**	10.42
mr	0.20	0.60	0.0	0.0	1.80	**3.11**
pt	8.90	3.20	13.10	0.20	**17.52**	13.73
ru	2.80	0.40	0	0.0	6.01	**6.51**
sr	3.20	2.10	1.70	0.50	**3.30**	2.0
ta	0.80	0.20	0.0	0.10	1.40	**3.11**
uk	2.30	0.30	0.0	0.10	5.21	**5.51**
vi	10.10	4.60	15.40	0.20	**28.43**	20.14
zh	1.0	0.60	0.0	0.0	**6.11**	5.41

Table 14: Performance of VEEF-Multi-SFT-8B and VEEF-Multi-DPO-8B models compared to Llama-2-7B, Mistral-7B, Bloom-7B, and LLaMAX2-7B models on XL-Sum (0-shot).

Language	SCORE	Llama2-7B	Mistral-7B	BLOOM Z-7B	LLaMAX2-7B	VEEF-Multi-SFT-8B	VEEF-Multi-DPO-8B
EN→RO	BLEU	17.18	13.66	1.88	24.52	26.32	**26.53**
	CHRF	44.20	41.47	20.09	51.94	54.23	**55.05**
RO→EN	BLEU	31.43	24.58	11.35	**36.02**	27.21	30.75
	CHRF	58.0	53.04	36.22	**60.85**	55.18	59.24
EN→DE	BLEU	20.01	19.41	3.76	26.31	**27.97**	26.70
	CHRF	48.31	49.25	23.27	53.95	**57.81**	57.54
DE→EN	BLEU	35.41	30.19	22.30	**37.05**	33.02	32.21
	CHRF	60.78	58.27	46.69	**61.90**	60.06	60.38
EN→FR	BLEU	24.97	24.24	17.73	32.86	**34.09**	33.22
	CHRF	52.34	52.08	41.02	59.53	**60.80**	60.78
FR→EN	BLEU	34.49	31.40	31.07	**36.00**	28.86	31.08
	CHRF	60.89	59.50	56.03	**61.64**	57.92	59.94
AR→EN	BLEU	12.51	9.13	**25.25**	29.76	21.44	22.88
	CHRF	36.18	32.64	47.64	52.68	42.95	**49.40**
EN→AR	BLEU	1.15	0.31	4.58	**10.47**	8.20	8.49
	CHRF	17.73	13.31	25.05	**40.27**	35.71	36.89
ZH→BO	BLEU	16.23	13.212	19.0	18.40	36.20	**40.60**
	CHRF	32.46	26.42	38.0	36.80	72.40	**81.20**
BO→ZH	BLEU	29.99	28.21	38.10	30.10	85.50	**86.80**
	CHRF	30.05	28.49	38.37	30.46	85.91	**87.21**
ZH→KK	BLEU	10.23	12.16	16.70	32.90	43.20	**44.80**
	CHRF	10.24	12.18	16.72	33.13	43.54	**44.88**
KK→ZH	BLEU	48.32	55.23	69.60	80.80	79.20	**83.0**
	CHRF	48.73	55.33	70.30	81.72	79.77	**83.29**
ZH→KO	BLEU	27.92	30.10	43.20	43.80	**44.90**	44.80
	CHRF	27.94	30.39	43.34	44.11	45.07	**45.20**
KO→ZH	BLEU	46.32	56.12	88.30	88.90	89.00	**90.30**
	CHRF	46.57	56.33	88.79	89.50	89.98	**91.07**
ZH→MN	BLEU	13.65	10.21	15.20	22.20	39.10	**41.60**
	CHRF	13.68	10.30	15.35	22.40	39.23	**41.94**
MN→ZH	BLEU	21.65	20.10	35.10	30.80	77.30	**79.40**
	CHRF	21.77	20.19	35.48	31.09	77.77	**79.47**
ZH→UG	BLEU	13.23	15.44	23.30	27.80	**33.40**	28.80
	CHRF	13.36	15.57	23.57	27.90	**33.77**	28.84
UG→ZH	BLEU	38.32	49.91	77.50	85.10	77.40	**78.80**
	CHRF	38.61	50.31	78.30	85.44	78.27	**79.62**
ZH→YUE	BLEU	40.34	43.12	58.30	60.0	59.70	**59.80**
	CHRF	40.49	43.26	58.53	60.07	59.77	**60.11**
YUE→ZH	BLEU	43.09	67.12	87.80	87.50	86.90	**87.90**
	CHRF	43.29	67.41	87.95	88.19	87.78	**88.15**

Table 15: Performance of VEEF-Multi-SFT-8B and VEEF-Multi-DPO-8B models compared to Llama-2-7B, Mistral-7B, Bloom-7B, and LLaMAX2-7B models on WMT14, WMT16,IWSLT2017,QHNU-test-tizh-CWMT2018, IMU-dev-mnzh-CWMT2017, WCM, and CWMT2018-TestSet-UC (BLEU, 0-shot).