IMPROVING MULTIMODAL LARGE LANGUAGE MODELS IN LOW-RESOURCE LANGUAGE CONTEXTS

Yufei Gao [1,2], **Jiaying Fei** [1], **Guohang Yan** [1*], **Yunshi Lan** [2*]
[1]Shanghai Artificial Intelligence Laboratory
[2]East China Normal University
[*]Corresponding author
{gaoyufei, feijiaying, yanguohang}@pjlab.org.cn
yslan@dase.ecnu.edu.cn

ABSTRACT

In recent years, open-source Multimodal Large Language Models (MLLM) have developed rapidly, but their strengths remain primarily in mainstream languages such as English and Chinese. Due to the relative scarcity of data for non-mainstream languages, these models perform poorly in low-resource languages, struggling not only to understand and generate them fluently but also to grasp the knowledge familiar to their speakers. Recognizing the importance of low-resource language data, this paper collects multimodal data containing small-language knowledge from relevant websites. Moreover, we propose a two-stage training approach to improving multimodal large language models in low-resource language contexts. In the first stage, multimodal capabilities are transferred to low-resource languages, while the second stage further supplements the model with the knowledge in the collected dataset. Experimental results demonstrate that this data collection strategy and training method effectively extend MLLM's multimodal capabilities to low-resource languages and enable multimodal large models to perform better in such contexts.

1 INTRODUCTION

Following the release of GPT-4V, open-source communities have actively responded by launching a series of competitive multimodal large language models (MLLMs), continuing to push the boundaries of multimodal AI research (Yin et al., 2024; Zhu et al., 2024; Contributors, 2023). However, open-source models that demonstrate strong performance in high-resource languages often fail to achieve comparable results in low-resource languages, primarily due to the imbalanced distribution of training data across different linguistic resources. This disparity may result in limited access to state-of-the-art artificial intelligence services for users of low-resource languages, potentially creating a technological divide in AI accessibility.

To promote the widespread adoption of AI services across diverse countries and regions, it is crucial to enhance the capabilities of multimodal large language models in low-resource language contexts. To resolve this problem, we propose a two-stage method. The language adaptation stage employs a distillation approach, using English as an intermediary language, to extend the image comprehension capabilities of InternVL2 to eight languages, including Arabic(AR), Korean(KO), Vietnamese(VI), Thai(TH), and Russian(RU), among others. In the knowledge enhancement stage, the knowledge of large models in low-resource languages is significantly enhanced by collecting familiar web-based multimodal data from low-resource language communities. Experimental results demonstrate that InternVL2 exhibits superior multimodal performance in low-resource languages after the language adaptation stage, and its knowledge scope is significantly expanded after sequential processing through both the language adaptation stage and the knowledge enhancement stage. Furthermore, the experiments validate the necessity of the language adaptation stage, confirming that this phase cannot be omitted.

2 RELATED WORK

2.1 MULTILINGUAL MULTI-MODAL LARGE LANGUAGE MODEL

Open-source large multimodal models have recently seen vigorous development, but their support for multilingual capabilities is insufficient. Some multimodal large models do not have a dedicated component for low-resource languages, such as Qwen-VL (Bai et al., 2023) and LLaVA (Liu et al., 2023). Others support only a select few languages or provide inadequate support, like InternVL2 Chen et al. (2024b), which supports only Chinese and English, and LLaVA-1.5, which has learned to follow Chinese instructions from fine-tuning on multilingual instructions without images. In contrast, closed-source large models exhibit greater multilingual support, such as Gmini (Team et al., 2023). Based on these observations, this paper aims to endow open-source multimodal large models with capabilities in other languages at a low cost, easily transferring multimodal abilities to other languages.

2.2 KNOWLEDGE DISTILLATION

The concept of knowledge distillation was first proposed by Hinton et al. (2015), with the core objective of compressing model size while maintaining performance, thereby facilitating model inference and deployment. Specifically, knowledge distillation achieves this by transferring knowledge from a more capable teacher model to a less capable student model. In the era of large models, the application of knowledge distillation extends beyond model lightweighting and is also used solely to enhance model capabilities (Xu et al., 2024). Research has shown that many popular large models incorporate knowledge distillation techniques (Lee et al., 2025). Additionally, self-distillation, introduced by Zhang et al. (2019), revolves around the idea of a model acting as its own teacher, where one part of the model guides the training of another part, or one training phase guides another.

In our approach, we combine knowledge distillation and self-distillation techniques. Notably, before transferring knowledge from the teacher model to the student model, we first perform a translation process on the knowledge generated by the teacher model.

2.3 LOW-RESOURCE LANGUAGES TRANSFER

With the flourishing development of natural language processing technology, how to transfer the capabilities of models to low-resource languages has garnered attention from researchers (Andersland, 2024; Yong et al., 2024; Zhang et al., 2024; Lai et al., 2024; Carlsson et al., 2022). To address the scarcity of data, researchers have proposed various efficient methods for generating high-quality data. The majority of these generation methods are related to machine translation, such as LexC-Gen (Yong et al., 2024), which uses a bilingual lexicon for word-to-word translation; Lai et al. (2024) and Andersland (2024) have translated existing datasets. In contrast, our approach not only teaches the model to use other languages but also imparts knowledge familiar to speakers of low-resource languages, enabling the model to truly "integrate" into the low-resource language community and achieve a deeper level of language understanding and application.

3 METHOD

3.1 LANGUAGE ADAPTATION STAGE

At present, the vast majority of open-source large multimodal models are unable to effectively perform multimodal tasks in low-resource languages, thereby limiting their benefits to a broader audience. To address the deficiency in small-language capabilities caused by the lack of small-language data, we referred to the Self-Distillation from Resource-Rich Languages method proposed in Zhang et al. (2024) and adapted it to multimodal models.

Distillation and Self-distillation Consider a teacher model $\mathcal{M}(\theta^{\mathrm{t}})$ parameterized by θ^{t} and a student model $\mathcal{M}(\theta^{s})$ parameterized by θ^{s}. Each data point in the multimodal dataset $D = \{(I_i, x_i), y_i\}_{i=1}^{N}$ has been labeled by the teacher model $\mathcal{M}(\theta^{\mathrm{t}})$. The input (I_i, x_i) goes through inference with the student model to obtain the output y_i^s. Distillation aims to minimize the cross-entropy (CE) loss between y_i^s and y_i. That is,

$$l_{CE}(y_i, y_i^s) = -\frac{1}{M}\sum_{m=1}^{M} \log P_{i_m}^s(y_{i_m}) \tag{1}$$

where M denotes the length of y_i, m represents the time step in the sequence, and $P_{i_m}^s$ is the probability distribution function of y_i^s at time step m. Thus, the overall distillation loss function can be expressed as

$$L_{distill} = \frac{1}{N}\sum_{i=1}^{N} l_{CE}(y_i^t, y_i^s) \tag{2}$$

We leverage self-distillation to efficiently and cost-effectively transfer multimodal capabilities to low-resource languages. Compared with traditional distillation, the advantage of self-distillation is that the teacher model and the student model are the same, allowing the student model to better adapt to the output patterns of the teacher model, thereby facilitating more effective learning.

When performing self-distillation, we set θ^t equal to θ^s, that is, after annotation by the model, the annotations are used for training the same model.

Language Adaptation We use the language that $\mathcal{M}(\theta^{\mathrm{t}})$ is most proficient in to annotate pure images, and then employ machine translation to translate the annotations into the target language. The above formula can be written as:

$$l_{CE}(y_i^{Trans}, y_i^{s,Trans}) = -\frac{1}{M}\sum_{m=1}^{M} logP_{i_m}^{s,Trans}(y_{i_m}^{Trans}) \tag{3}$$

$$L_{LA} = \frac{1}{N}\sum_{i=1}^{N} l_{CE}(y_i^{Trans}, y_i^{s,Trans}) \tag{4}$$

where the superscript $Trans$ denotes the translated content.

3.2 Knowledge Enhancement Stage

Although multimodal large models have learned to use low-resource languages, as shown in Figure 2, models that have only undergone the Language Adaptation Stage still cannot recognize political figures frequently appearing on Arabic websites. In other words, merely mastering a low-resource language is not enough for the model to "integrate" the knowledge of that language into use. To this end, we have designed the Knowledge Enhancement Stage to immerse the model in the local language environment for learning.

Data Collection To expand the knowledge base of our model, we crawled image-text pairs from news websites in each respective language. To ensure data quality, we conducted a thorough cleaning and filtering process for the images, which included removing duplicates, eliminating irrelevant image-text pairs, and excluding inappropriate content. Ultimately, we amassed nearly one hundred thousand image-text pairs for each language. We denote this dataset as $D_{KE} = \{((I^{KE}, x^{KE}), y^{KE})\}$, where I^{KE} represents the image, x^{KE} represents a random instruction from manually written instruction pool and y^{KE} represents the text.

Knowledge Enhancement We employ the identical loss function as utilized in the Language Adaptation Stage, specifically the cross-entropy loss, as the optimization objective. Denote the model trained in the previous stage as $\mathcal{M}'$, the mathematical formulation of the loss function is presented as follows:

$$L_{KE} = \frac{1}{N}\sum_{i=1}^{N} l_{CE}(y_i^{KE}, \mathcal{M}'(I_i^{KE}, x^{KE})) \tag{5}$$

where $\mathcal{M}'(I_i^{KE}, x^{KE})$ represents the inference result of the model trained in the previous stage.

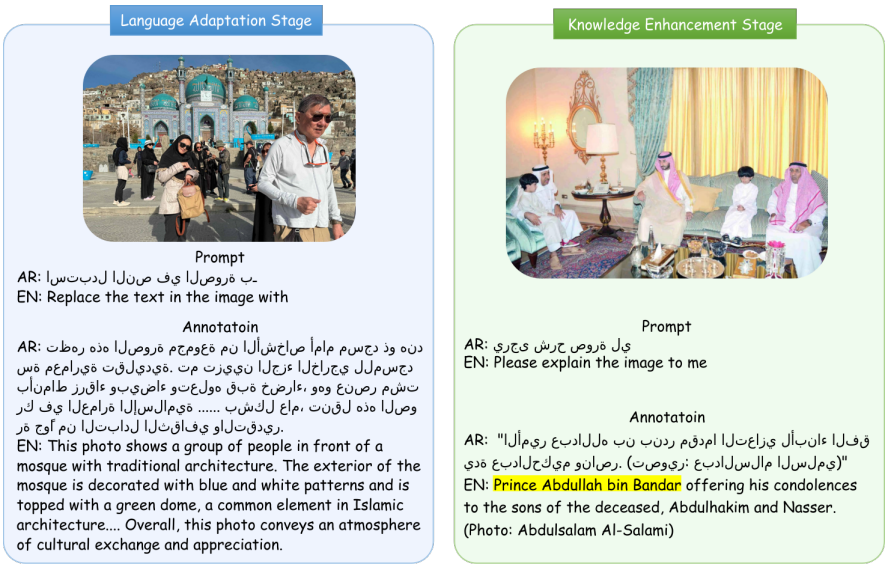

Figure 1: Representative Data Samples from Both Training Stages

3.3 Overview

The first stage aims to enable the model to perform general multimodal tasks using low-resource languages. However, the model trained in the first stage is not sufficiently capable of answering questions related to the knowledge of low-resource languages. Therefore, a second stage is necessary. The data for the second stage is sourced from websites in the corresponding low-resource languages, which contain rich knowledge specific to these languages.

4 Experiments

4.1 Setup

We utilize InternVL2-8B (Chen et al., 2024a) as our foundational model. During the language adaptation stage, considering the model's superior performance in Chinese, we employ Chinese as a bridge language to facilitate the learning of eight low-resource languages: Arabic (AR), Czech (CS), Hungarian (HU), Korean (KO), Russian (RU), Serbian (SR), Thai (TH), and Vietnamese (VI). The image captioning task is strategically selected as the primary learning framework for these languages. To address concerns of over-fitting, we also manually crafted 20 distinct instructions, which are randomly selected and used during the training process to enhance the model's generalization capabilities. As illustrated in Figure 1, each training instance comprises an image-instruction pair with its corresponding ground-truth caption. More details of the instruction templates and training dataset specifications can be found in Appendix A.1.

Implementation Details Our code is implemented based on DeepSpeed (Rasley et al., 2020) on two NVIDIA A100-SXM4-80GB GPUs. We set the training duration to 1 or 2 epochs with an automatically calculated learning rate. Other hyperparameters are set according to InternVL2 (Chen et al., 2024a).

Metrics

- **Meteor**: Measures translation quality by aligning and matching words, considering synonyms and stemming (Denkowski & Lavie, 2014).
- **Bleu**: Evaluates precision of n-gram overlaps between generated and reference texts (Papineni et al., 2002).
- **Rouge-L**: Assesses recall-based similarity by measuring the longest common subsequence (Lin, 2004).

4.2 MAIN RESULTS

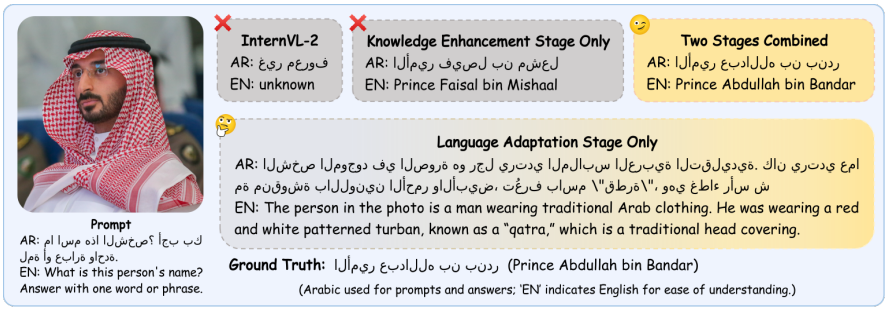

Figure 2: A case study demonstrates the performance evolution of the InternVL-8B model in Arabic dialogue tasks. Taking the recognition of the prominent Arab figure, Prince Abdullah bin Bandar, as an example: (1) The baseline model fails to identify the individual; (2) After training solely with the Knowledge Enhancement Stage, the model exhibits character confusion; (3) With only the Language Adaptation Stage training, the model generates only superficial image descriptions; (4) Following complete two-stage training, the model achieves accurate person recognition.

4.2.1 LANGUAGE ADAPTATION STAGE RESULTS

	AR	SR	RU	CS	KO	TH	VI	HU
				Meteor				
InternVL2-8B	26.07	2.7	7.71	3.37	14.54	19.95	18.19	0.11
Qwen2-VL-7B-Instruct	15.49	2.33	6.54	6.03	12.93	17.14	16.77	6.37
Phi-3.5-vision-instruct	20.94	0.36	3.34	3.67	2.04	8.46	5.95	2.67
InternVL2-8B-trained	**37.9**	**17.29**	**14.81**	**15.59**	**29.39**	**35.10**	**33.41**	**16.16**
				Bleu-1				
InternVL2-8B	5.75	4.07	14.44	3.84	7.19	0.35	16.71	0.14
Qwen2-VL-7B-Instruct	5.83	1.94	7.26	5.42	4.42	0.82	15.47	9.48
Phi-3.5-vision-instruct	4.26	0.33	4.11	4.24	0.98	0.22	2.95	5.12
InternVL2-8B-trained	**35.35**	**37.88**	**32.72**	**32.38**	**27.78**	**3.17**	**52.58**	**32.9**
				Bleu-2				
InternVL2-8B	1.27	0.14	5.19	1.07	2.52	0.17	7.13	0.04
Qwen2-VL-7B-Instruct	2.49	0.33	3.31	2.41	1.23	0.3	9.02	3.2
Phi-3.5-vision-instruct	1.37	2.57e-8	0.98	1.04	0.29	0.07	0.65	0.73
InternVL2-8B-trained	**21.96**	**21.31**	**18.44**	**19.25**	**16.56**	**0.81**	**37.83**	**18.14**
				Bleu-3				
InternVL2-8B	0.1	0	1.97	0.24	0.49	0.06	2.89	2.33e-5
Qwen2-VL-7B-Instruct	1.09	0.09	1.68	1.18	0.19	0.14	5.27	0.77
Phi-3.5-vision-instruct	0.37	2.65e-13	0.22	0.37	0.03	5.12e-6	0.08	9.40e-5
InternVL2-8B-trained	**14.51**	**13.81**	**12.09**	**12.92**	**9.9**	**0.07**	**27.74**	**11.28**
				Bleu-4				
InternVL2-8B	0.03	0	0.65	0.07	0.04	7.38e-8	0.9	3.03e-9
Qwen2-VL-7B-Instruct	0.39	0.02	0.71	0.47	0.07	**1.66e-7**	2.93	0.15
Phi-3.5-vision-instruct	0.17	4.25e-15	0.07	0.08	3.23e-8	7.57e-10	8.06e-8	1.1e-3
InternVL2-8B-trained	**9.99**	**9.16**	**7.97**	**8.98**	**5.61**	9.16e-8	**20.96**	**7.01**
				Rouge-L				
InternVL2-8B	5.23	6.41	12.73	6.25	6.25	0.5	12.39	0.22
Qwen2-VL-7B-Instruct	11.3	5.5	12.86	1.11	7.85	1.31	16.84	11.3
Phi-3.5-vision-instruct	5.4	1.01	6.78	7.04	1.07	0.42	3.21	5.57
InternVL2-8B-trained	**24.23**	**23.74**	**21.62**	**22.73**	**19.71**	**3.38**	**30.61**	**23.22**

Table 1: Performance Evaluation of the Language Adaptation Stage

Table 1 shows the experimental results of the image caption task. From the experimental results, we can observe that:

(1) The efficacy of Language Adaptation Stage is pronounced. Across a suite of evaluation metrics, we note substantial enhancements, with particularly standout performances observed in Vietnamese, Arabic, and Serbian datasets.

(2)The efficiency of the Language Adaptation Stage is also commendable. For the majority of the multilingual caption datasets, the InternVL2-8B model was able to achieve the aforementioned

outcomes after a single training epoch with under 90,000 data samples, including a mere 40,000 samples for Russian. Notably, only the Hungarian dataset required more than 80,000 samples for two training epochs. Critically, the realization of these results was achieved without any expenditure on human-annotated data.

Question	**InternVL2-8B Two Stages Combined**	**InternVL2-8B Knowledge Enhancement Stage Only**	**Annotation**
What is the name of this person? Answer with one word or phrase.	Prince Abdullah bin Bandar	Prince Faisal bin Mishal	Prince Abdullah bin Bandar
What is this city? Answer with one word or phrase.	The royal city of Riyadh celebrates the twelfth Saudi National Day from the date.	The International City of Riyadh	Riyadh
Describe this picture in Arabic, about 10 words.	Messi cries after winning over the Argentine team in the first goal of the 2018 World Cup.	Messi suffers from the high injury that causes the high injury that causes the high injury that causes the high injury that causes the high injury that causes the high injury	Argentina was defeated and Messi was very frustrated.
Where is this? Answer with one word or phrase.	Sudanese settlers carry water for cleaning in the old city of Riyadh in Sudan. (AFP)	The ceremony in which it was filmed in the old city of Sudan detained for two years causing the civil war in which it was filmed in the old city of Sudan	The Great Mosque in Geneva
Where is this? Answer with one word or phrase.	The Prophet's Mosque in Riyadh	The Grand Mosque in Mecca	The Great Mosque or the Prophet's Mosque
What is this event? Answer in a short sentence.	The Ukrainian forces prepare to fight the Russian in the explosion site that disrupted the Ukrainian city near the Russian border	The Ukrainian forces prepare for victory after the explosion that led to the collapse of the central hospital in the occupied eastern Ukrainian city between the Russian forces and the citizens	Ukraine faces a Russian attack.
What is the name of this person? Answer with one word or phrase.	Michel Assad	Michel Alouites Prince Michel Alouites the First	Michel Aoun
What is the name of this person? Answer with one word or phrase.	Prince Abdullah bin Mohammed bin Abdulaziz	King Salman bin Abdulaziz	Salman bin Abdulaziz Al Saud
What is the name of this tower? Answer with one word or phrase.	Burj Burkhalin	Burj Belgical	Burj Khalifa
Which holiday is this? Answer with one word or phrase.	Children working in a grain factory in Sudan	Children working in a household materials factory in Sudan.	Day of Plastering
What is the name of this tower? Answer with one word or phrase.	Burj Beit M	Burj Belgium	Burj Al Arab

Table 2: Performance of Knowledge Enhancement Stage.(Translated from Arabic.)

4.2.2 KNOWLEDGE ENHANCEMENT STAGE RESULTS

As demonstrated in Table 2, a comparative analysis between the fully trained two-stage model and the model trained solely with the second stage reveals that the two-stage learning framework not only enables Arabic language proficiency but also facilitates the acquisition of Arabic-specific knowledge.

5 CONCLUSION AND FUTURE WORK

This paper proposes an innovative solution based on previous research, aiming to address the dual challenges of low-resource language data and knowledge scarcity. Our approach employs a two-stage training framework: first, the multimodal large model acquires low-resource language capabilities through a language adaptation stage. Then, high-quality image-text pair data is crawled from local websites in the target language to construct the training set for the knowledge enhancement stage, thereby systematically completing the model's knowledge system in a specific cultural

context. Experimental results show that the language adaptation stage demonstrates significant effectiveness and computational efficiency in improving the model's language capabilities, while the knowledge enhancement stage successfully integrates unique knowledge from the Arabic cultural sphere into the model.

In the future, we will incorporate more baselines and ablation studies to validate our approach on a broader scale. For the samples in the current experimental results that are still unsatisfactory, we will identify the underlying causes and further optimize the details of the fine-tuning process.

6 LIMITATIONS

Despite the two-stage training process, the model occasionally fails to fully acquire the knowledge embedded in the training data. This limitation may stem from two primary factors: firstly, the inherent inconsistency between news images and their corresponding headlines in web sources, which could potentially mislead the learning process; secondly, the current training paradigm might require more sophisticated design, particularly in implementing a progressive adjustment strategy that gradually increases the proportion of knowledge-based data while correspondingly decreasing the proportion of linguistic competence data throughout the training phases.

ACKNOWLEDGMENTS

The research was supported by Shanghai Artificial Intelligence Laboratory, the National Key R&D Program of China (Grant No. 2022ZD0160104), and the Science and Technology Commission of Shanghai Municipality (Grant No. 22DZ1100102).

REFERENCES

Michael Andersland. Amharic llama and llava: Multimodal llms for low resource languages, 2024. URL https://arxiv.org/abs/2403.06354.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. Cross-lingual and multilingual CLIP. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 6848–6854, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.739/.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, and Lucia Specia (eds.), *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 376–380, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3348. URL https://aclanthology.org/W14-3348/.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. Llms beyond english: Scaling the multilingual capability of llms with cross-lingual feedback, 2024. URL https://arxiv.org/abs/2406.01771.

Sunbowen Lee, Junting Zhou, Chang Ao, Kaige Li, Xinrun Du, Sirui He, Jiaheng Liu, Min Yang, Zhoufutu Wen, and Shiwen Ni. Distillation quantification for large language models, 2025. URL https://arxiv.org/abs/2501.12619.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 3505–3506, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3406703. URL https://doi.org/10.1145/3394486.3406703.

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

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models, 2024. URL https://arxiv.org/abs/2402.13116.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, November 2024. ISSN 2053-714X. doi: 10.1093/nsr/nwae403. URL http://dx.doi.org/10.1093/nsr/nwae403.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. Lexc-gen: Generating data for extremely low-resource languages with large language models and bilingual lexicons, 2024. URL https://arxiv.org/abs/2402.14086.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation, 2019. URL https://arxiv.org/abs/1905.08094.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11189–11204, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.603. URL https://aclanthology.org/2024.acl-long.603/.

Shaolin Zhu, Supryadi, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, and Deyi Xiong. Multilingual large language models: A systematic survey, 2024. URL https://arxiv.org/abs/2411.11072.

A APPENDIX

A.1 TRAINING DATA INFORMATION

	Train Set Size	Test Set Size
AR	80,000	100
CS	90,000	100
HU	82,849	100
KO	77,202	100
RU	28,504	100
SR	90,000	100
TH	49,999	100
VI	85,000	100

Table 3: Language Adaptation Stage Dataset Size information.

	Train Set Size
AR	80,000

Table 4: Knowledge Enhancement Stage Dataset Size information. The evaluation process is currently conducted through manual assessment by human experts.

AR	EN
استبدل النص في الصورة بـ	Replace the text in the image with
استبدل النص في الصورة بـ	The replacement text for this image is
تم استبدال الصورة بـ	Due to privacy issues, the image has been replaced with
يمكن استبدال الصورة بـ	The image can be replaced with
نص الصورة البديل هو	The alternative text for this image is
استبدل الصورة بالنص	Replace the image with text
استبدل الصورة بالنص	The result of replacing the image with text is
يرجى استبدال الصورة بالنص	Please replace the image with text
ما هو وصف الصورة؟	What is the text description of this image?
صف الصورة بـ	The description of the image is
وصف الصورة هو	The description of this image is
صف الصورة	Please describe this image
أنا شخص مكفوف، يرجى وصف الصورة لي	I am visually impaired, please describe this image to me
أنت الآن معلم، يرجى شرح الصورة للطلاب	You are now a teacher, please explain this image to the students
أنت مقدم البرنامج، يرجى شرح الصورة للجمهور	You are a presenter, please explain this image to the audience
يرجى شرح الصورة لي	Please explain the image to me
يرجى شرح صورة لي	Please describe an image to me
صف الصورة	Describe this image
ماذا يوجد في الصورة؟	What is in the image?
ما هو محتوى الصورة؟	What is the content of the image?
قدم الصورة	Introduce this image

Table 5: 20 manually crafted distinct instructions.