🧘 CALM: Unleashing the Cross-Lingual Self-Aligning Ability of Language Model Question Answering

Yumeng Wang[2] **Zhiyuan Fan**[2] **Qingyun Wang**[1] **Yi R. (May) Fung**[2*] **Heng Ji**[1*]
[1]University of Illinois Urbana-Champaign, [2]HKUST
ywanglu@connect.ust.hk yrfung@ust.hk hengji@illinois.edu

Abstract

Large Language Models (LLMs) are pretrained on extensive multilingual corpora to acquire both language-specific cultural knowledge and general knowledge. Ideally, while LLMs should provide consistent responses to culture-independent questions across languages, we observe significant performance disparities. To address this, we explore the **C**ross-Lingual Self-**A**ligning ability of **L**anguage **M**odels (**CALM**) to align knowledge across languages. Specifically, for a given question, we sample multiple responses across different languages, and select the most self-consistent response as the target, leaving the remaining responses as negative examples. We then employ direct preference optimization (DPO) to align the model's knowledge across different languages. Evaluations on the MEDQA and X-CSQA datasets demonstrate CALM's effectiveness in enhancing cross-lingual knowledge question answering, both in zero-shot and retrieval-augmented settings. We also found that increasing the number of languages involved in CALM training leads to higher accuracy and consistency. We offer a qualitative analysis of how cross-lingual consistency can enhance knowledge alignment and explore the method's generalizability[1].

1 Introduction

LLMs have been pre-trained on various knowledge domains in multiple languages, capturing extensive world knowledge (Yu et al., 2024). This knowledge can be either sociocultural-dependent (Sun et al., 2023; Liu et al., 2025) or sociocultural-independent (Tang et al., 2024; Huang et al., 2024a). Ideally, LLMs should deliver consistent responses to the sociocultural-independent questions. However, due to the imbalance of the pretraining data, such knowledge is not well-aligned (Qi et al., 2023; Xu et al., 2024; Wu et al., 2025a). Research indicates that LLMs exhibit varying proficiency when addressing the same task across different languages (Xu et al., 2024; Huang et al., 2024b). This variability stems from the difficulty of accessing knowledge encoded in one language while using others.

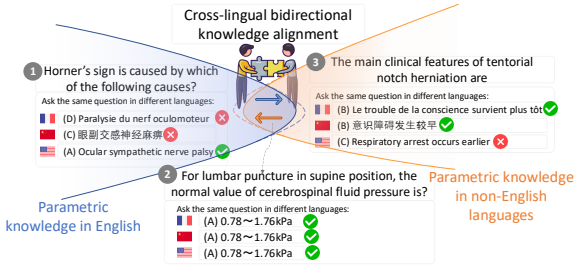

Figure 1: Knowledge is not well-aligned across languages. (1) represents knowledge encoded in English that is difficult to retrieve from other languages. (2) is the knowledge that is already well-aligned across languages. (3) is the knowledge encoded in other languages that is difficult to retrieve in English. Ideally, we want all the culture-independent knowledge to fall into (2).

To bridge the gap, recent papers introduced cross-lingual consistency (Qi et al., 2023), which pertains to the capacity to provide consistent responses across different languages when presented with the same query. The ultimate goal is to achieve language-agnostic question-answering proficiency in LLMs, enabling them to generalize effectively in multilingual environments. Gao et al. (2024) highlighted the positive impact of multilingual pretraining and instruction tuning on enhancing cross-lingual consistency. However, it also pointed out that current LLMs still face challenges in scaling up to improve cross-lingual knowledge retrieval capabilities. Chen et al. (2023) utilized translation to develop a multilingual math reasoning instruction dataset. However, the challenge lies in the labor-intensive nature of obtaining high-quality translations and annotating data. She et al. (2024) lever-

* Corresponding author.
[1]The source code and data of this paper is available on https://github.com/wangym2/CALM.

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 2809–2817

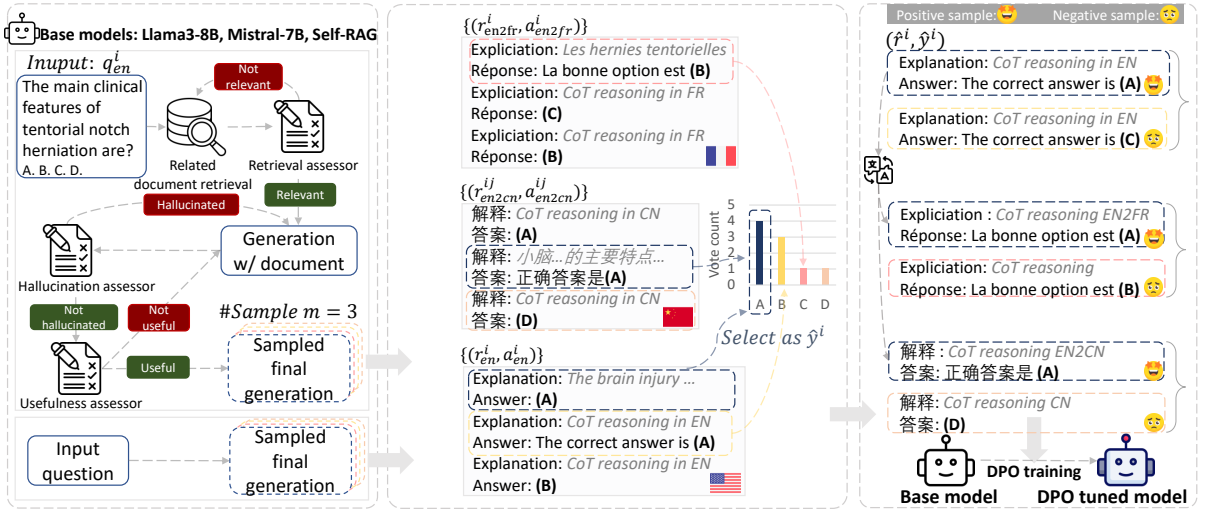

Figure 2: An example of the three stages in our proposed method assuming a question input originally in English.

aged translation consistency as a reward model to align the reasoning processes in other languages with the dominant language. Nevertheless, this approach may diminish the diversity of knowledge or reasoning introduced by different languages. Huang et al. (2024b) enhanced the multilingual culture commonsense reasoning by implementing a multi-agent framework to aggregate the knowledge from diverse languages. In this work, we focus on leveraging multilingual knowledge aggregation by adopting preference optimization for model tuning.

To address the challenges of (1) establishing a scalable framework for aligning culture-independent knowledge across different languages and (2) lacking high-quality annotated data for training, we propose CALM, a method that encourages consistent answers to the same questions in different languages, motivated by the observation (Figure 1) that non-English languages often contain complementary knowledge missing in English outputs. In Figure 3, majority-voted answers consistently outperform English-only responses, making them viable alignment targets despite occasional factual inaccuracies. Exclusively aligning all other languages to English fails to leverage the LLM's full multilingual knowledge potential, whereas CALM's language-agnostic voting mechanism synthesizes cross-lingual insights.

Our approach leverages direct preference optimization (DPO) (Rafailov et al., 2024) to facilitate cross-lingual alignment. The approach involves three steps. First, we sample a variety of multilingual Chain-of-Thought (CoT) outputs from the models. Next, we conduct majority voting on the sampled outputs in different languages, selecting the answer with the highest vote as positive. Finally, we pair the positive sample with all other answers that are inconsistent with it, utilizing these pairs for DPO training. Moreover, we expand this framework to integrate external knowledge by combining Self-supervised Retrieval-Augmented Generation (Self-RAG) (Asai et al., 2023) with DPO.

We conduct experiments on the challenging MEDQA (Jin et al., 2020) and the multilingual X-CSQA (Lin et al., 2021) datasets, each representing general knowledge and commonsense knowledge. On average, CALM boosts the accuracy on MEDQA and X-CSQA by +3.76% and +5.55% respectively. Our key contributions are summarized:

- We propose CALM, a label-free approach to effectively align the culture-independent knowledge by encouraging cross-lingual consistency, enabling the model to enhance its knowledge accuracy and consistency (Huang et al., 2023).
- We conduct experiments in both zero-shot Chain-of-Thought and retrieval augmented settings, utilizing Llama3-8B-Instruct (Dubey et al., 2024), Self-RAG (Asai et al., 2023), and Mistral-7B-Instruct-v0.2 (Jiang et al., 2023). The outcomes highlight the efficacy of our approach in aligning internal and external knowledge.
- We further evaluate the cross-language and cross-dataset generalizability of CALM, showcasing its robustness and scalability.

Model	MEDQA (%)			X-CSQA (%)								
	EN	ZH	ACC_{avg}	EN	ZH	FR	IT	DE	JA	ACC_{avg}	Consis	AC3
Llama	60.1	56.2	58.2	73.1	52.1	60.8	59.8	57.5	49.2	62.0	57.73	58.24
+ SFT	62.4	57.1	59.8	73.8	53.2	62.3	60.0	59.8	51.0	63.1	59.67	60.82
+ **CALM**	**63.5**	**59.5**	**61.5**	**74.1**	**57.6**	**65.0**	**64.7**	**60.9**	**53.6**	**64.8**	**61.13**	**61.70**
Self-RAG	62.6	57.1	59.9	-	-	-	-	-	-	-	-	-
+ SFT	63.8	60.3	62.1	-	-	-	-	-	-	-	-	-
+ **CALM**	**64.7**	**63.7**	**64.2**	-	-	-	-	-	-	-	-	-
Mistral	49.8	36.4	43.1	60.1	48.3	51.6	50.7	49.4	43.0	53.3	50.51	50.51
+ SFT	50.3	37.9	44.1	67.7	48.8	53.7	56.6	55.6	44.1	56.7	53.27	53.83
+ **CALM**	**52.9**	**38.5**	**45.7**	**68.1**	**56.8**	**56.8**	**57.7**	**58.6**	**50.5**	**60.6**	**57.27**	**57.67**

Table 1: Model accuracy percentage score on the test set of MEDQA and X-CSQA in different languages. "ACC_{avg}" denotes the average traditional accuracy of all languages, which represents the overall level of domain knowledge of the model. The bold text represents the best result in the given model. Note that there are no X-CSQA results for Self-RAG because there are no documents available for retrieval. The full result of MEDQA can be found in Table 9.

2 Method

To encourage cross-lingual consistency, CALM samples a variety of Chain-of-Thought (CoT) (Wei et al., 2024; Kojima et al., 2024) responses from different languages, and leverages response consistency (Wang et al., 2023; Wu et al., 2025b) as the learning signal. By selecting the most voted response as the positive sample, we construct the preference pairs and adopt DPO to optimize the preference. As the winning response may be any language, we preserve the diverse knowledge from languages other than English. We verified our approach in a retrieval-augmented setting, showing that our approach boosts the multilingual transferability of both internal and external knowledge. The proposed framework is shown in Figure 2. Our method comprises multilingual response sampling, self-consistency-based preference pair construction, and multilingual knowledge alignment.

2.1 Multilingual response sampling

Translation For monolingual dataset, where a series of multiple choice questions are provided in its primary language (e.g., English), denoted as $Q_{en} = \{q^i_{en}\}^N_{i=1}$, we first translate them into two additional languages, say Chinese (Q_{en2cn}) and French (Q_{en2fr}). For multilingual datasets, this translation step is omitted, and the parallel questions in different languages are utilized directly.

CoT answer generation We apply multiple path decoding with temperature *T = 1* on each variant of the question q^i_* for all $i = 1, ..., N$ and * be any language in $\{en, en2fr, en2cn\}$ to generate m pairs of CoT explanations and answers $\{(r^{ij}_*, y^{ij}_*)\}^m_{j=1}$, where y denotes one of the predicted choice (A, B, C,...). The model is instructed to output an "Explanation" followed by an "Answer" to conform with the CoT format (Wei et al., 2024).

2.2 Self-consistency based preference pair construction

Self-consistency CALM assumes that the answer with the most votes reflects the highest model confidence (Xiong et al., 2024; Kabra et al., 2023), making it more likely to be correct (Wang et al., 2023). use majority voting to identify the most popular option $\hat{y}_i$ from all multilingual answers, though $\hat{y}_i$ may not necessarily match the ground truth answer. We designate the most self-consistent answer as the positive sample.

Preference pair After obtaining a set $S = \{(r^{ik}, y^{ik})\}_k$ of the most voted explanation-answer pair that satisfies $\forall y^{ik} \in \{(r^{ik}, y^{ik})\}_k, y^{ik} = \hat{y}^i$, we pair each of the positive samples with negative samples. Note that the positive samples are not necessarily in English. Hence, we aggregate the internal knowledge of both English and non-English languages. Negative samples are inconsistent with the positive ones, i.e., $y_{negative} \neq \hat{y}^i$. For each positive-negative sample pair, the positive sample is translated into the language of the negative sample. The final preference pairs of the i-th question are $p^i = \{p^i_w : (\hat{r}^i{}_{trans}, \hat{y}^i), p^i_l : (r^i, y^i)_{neg}\}$.

2.3 Multilingual knowledge alignment

We adopt DPO as the alignment approach using the preference pairs (p_w, p_l) obtained from 2.2, where p_w is preferred over p_l. Given an input question q, we optimize the following objective:

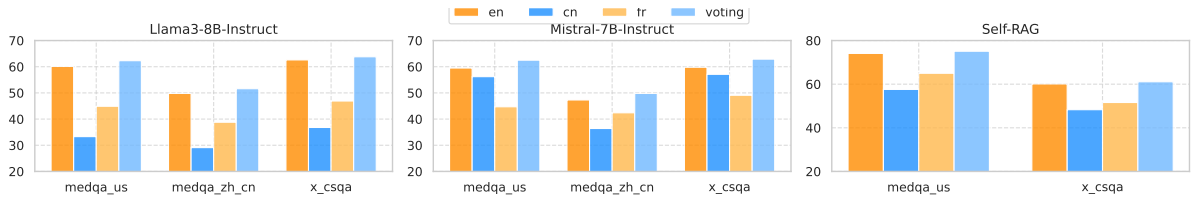

Figure 3: Visualization of mono-lingual (EN, ZH-CN, FR) percentage accuracy against the multilingual majority voting accuracy. The multilingual majority-voting result always has the highest accuracy. The proportion of each language in the CALM training data is in Table 7.

$$L_{\text{DPO}}(\pi_\theta; \pi_{ref}) = \mathbf{E}(q, p_w, p_l) \sim$$
$$\mathcal{D}\left[\log \sigma \left(\beta \log \frac{\pi_\theta(p_w|q)}{\pi_{\text{ref}}(p_w|q)} - \beta \log \frac{\pi_\theta(p_l|q)}{\pi_{\text{ref}}(p_l|q)}\right)\right]$$

3 Experiment and Results

3.1 Datasets and Metrics

We perform experiments on the following datasets:

- **MEDQA:** Zero-shot question answering, and Self-RAG's noisy evidence retrieval (Jin et al., 2020) over multiple evidence on medical multi-choice questions.
- **X-CSQA:** General multilingual commonsense question answering, including parallel questions from English, Chinese, French, Italian, German, and Japanese.

We adopt the multilingual consistency metrics introduced by (Wang et al., 2024; Lin et al., 2024), which includes *traditional accuracy*, *consistency* and *AC3*. Traditional accuracy refers to the accuracy of the multiple-choice questions. *Consistency* is intended to measure if the model delivers consistent responses to the same question in different languages. A higher consistency score implies that multilingual LLMs can provide consistent responses across languages, which is irrelevant to the accuracy. For datasets like X-CSQA that contains a set of questions $Q = \{q^i\}_{i=1}^N$ across six languages, the consistency metric is defined as:

$$M_{\{l_1,\ldots,l_s\}} = \frac{\sum_{i=1}^N 1\{y_i^{l_1} = y_i^{l_2} = \ldots = y_i^{l_s}\}}{N}$$

in which $y_i^{l_s}$ denotes the answer to the i-th multiple choice question given by language l_s. The final multilingual consistency is given by:

$$Consistency_s = \frac{\sum_{\{l_1,l_2,\ldots,l_s \in C(a,q_i)\}} M_{\{l_1,l_2,\ldots,l_s\}}}{C_6^s}$$

AC3 is a metric combining accuracy and cross-lingual consistency, which is more robust for this multilingual task. The formulation is given by:

$$AC3_s = 2 \times \frac{Accuracy \times Consistency_s}{Accuracy + Consistency_s}$$

By considering both accuracy and multilingual consistency, we can measure the knowledge gain and the cross-lingual consistency.

3.2 Baselines

Base models Our experiments utilize three base models, including Llama3-8B-Instruct (Dubey et al., 2024), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), and Self-RAG (Asai et al., 2023). The testing results from the first two models demonstrate the efficacy of our approach in aligning internal knowledge, while the result from the last model highlights its proficiency in aligning external knowledge. The primary baseline is the direct inference results from all base models.

Supervised finetuning on preferred samples To prove the necessity of DPO in training, we adopt supervised fine-tuning (SFT) (Luong and Manning, 2015) on preferred samples, namely using the most voted answers as SFT labels.

3.3 Results

In Table 1, CALM has encouraged the model to produce more accurate and consistent answers in all settings, outperforming the base model and the supervised fine-tuned model under all settings. Notably, the performance gain in X-CSQA surpasses that of MEDQA, which is likely due to the involvement of more languages participating, thereby activating more internal knowledge. Therefore, we can conclude that our approach has successfully facilitated the cross-lingual self-alignment.

4 Discussion

4.1 Accuracy of the positive samples

In Figure 3, we observe that the most self-consistent answer does not always align with the factually correct answer. Although the self-consistent answer's accuracy slightly surpasses monolingual accuracy, the improvement remains modest. This raises an important question regarding the effectiveness of noisy labels in CALM's training process. To better understand this phenomenon, we examine examples of the preference data generated by CALM in Table 4 in the Appendix. The example shows that, although the preferred data may be factually incorrect, it often demonstrates better context awareness, which can lead the model to generate more accurate answers.

Model	MEDQA		X-CSQA		
	EN	ZH	EN	ZH	FR
Llama3-SFT w/ GT	62.5	58.8	73.5	53.8	63.8
Llama3-DPO w/ GT	62.5	59.3	74.0	54.3	64.1
Self-RAG-SFT w/ GT	63.6	62.3	-	-	-
Self-RAG-DPO w/ GT	64.5	63.8	-	-	-
Mistral-SFT w/ GT	50.9	36.9	73.0	51.6	60.1
Mistral-DPO w/ GT	52.4	38.1	73.2	51.8	55.0

Table 2: Two additional baselines: DPO and SFT with ground truth. In this setting, we only keep the portion of DPO and SFT data that are factually correct.

4.2 SFT and DPO with ground truth

Using ground truth from X-CSQA and MEDQA, we evaluate supervised SFT and DPO, retaining only preference pairs and SFT data where positive samples match ground truth. In Table 2, supervised methods do not significantly outperform CALM, suggesting that guiding the model toward more confident and self-consistent answers can achieve comparable correctness even without ground truth.

Model	MEDQA			X-CSQA	
	EN	FR	ZH-CN	EN	ZH-CN
Llama3-8B	73.4	62.7	53.8	60.9	57.9
Mistral-7B	70.8	55.1	55.6	52.9	37.2

Table 3: We investigate the cross-dataset generalizability. The table shows the result of training on MEDQA and testing on X-CSQA, or training on X-CSQA and testing on MEDQA. Both settings surpass the baseline.

4.3 Generalizability

Cross-dataset generalizability To evaluate the generalizability, we conduct cross-dataset experiments by training models on X-CSQA and testing them on MEDQA, and vice versa. Table 3 reveals that while the out-of-domain accuracy falls below the in-domain accuracy, it consistently exceeds the in-domain performance of the SFT baseline. This underscores the capability of CALM-trained models to provide multilingually consistent answers, even when faced with unseen tasks or domains. These findings suggest that CALM enhances in-domain performance and fosters robustness across different types of domains.

Cross-lingual generalizability We implement CALM training sequentially, beginning with English and incrementally adding French and Chinese, progressing from high-resource to low-resource languages. At each step, we evaluate test accuracy across all languages. To assess CALM's effectiveness in untrained languages, we include Japanese, Italian, and German in the test set, none of which were included during training. In Table 10 in the Appendix, CALM demonstrates greater effectiveness as more languages participate in majority voting. Notably, even untrained languages exhibit accuracy improvements, suggesting that CALM's alignment mechanism fosters a unified understanding of knowledge across languages, thereby enhancing overall comprehension. This aligns with She et al. (2024), which similarly observe cross-lingual generalizability in multilingual reasoning tasks.

5 Conclusion

We introduce CALM, a novel framework to facilitate the alignment of LLM's knowledge across different languages. We observe that CALM is more effective when more languages are involved in the training, due to internal knowledge aggregation. Additionally, CALM outperforms ground truth DPO and SFT. It shows that although some of the positive samples are factually incorrect, they also contribute to the accuracy gain in CALM, possibly because more consistent answers often have better task understanding and can lead the model towards more correct answers. Through comprehensive experiments, we demonstrate the effectiveness of CALM in achieving robust cross-lingual knowledge alignment.

Limitations

One of the main limitations of our study is that due to the constraints of computational resources, we are unable to perform experiments on larger models. For the same reason, we are also not able to perform full-parameter fine-tuning and can only use LoRA DPO fine-tuning as an alternative. The translations in the experiment are done by Google Translate API, which may not be accurate sometimes because the dataset contains a many challenging medical terminology, hindering our final performance. For the DPO training data construction, since the accuracy after majority-voting is still low, the final alignment performance may be constrained by the noisy labels in the positive samples. Training one language after another can result in performance degradation in other languages. Future work can further investigate continual learning in multilingual knowledge alignment.

Ethics Statements

In this paper, we present a method to align knowledge across multiple languages, ensuring equitable access to LLMs for users from diverse linguistic backgrounds. Our approach utilizes the model's own outputs to perform cross-lingual alignment without the need for human annotations. By reducing dependence on manual labeling, this method enhances fairness, scalability, and inclusivity in multilingual AI, furthering the democratization of LLMs across global communities.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Yangqiu Song, Dongmei Zhang, and Jia Li. 2023. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *Preprint*, arXiv:2310.20246.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. *Preprint*, arXiv:2404.04659.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.

Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. 2024a. Do LVLMs understand charts? analyzing and correcting factual errors in chart captioning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 730–749, Bangkok, Thailand. Association for Computational Linguistics.

Yue Huang, Chenrui Fan, Yuan Li, Siyuan Wu, Tianyi Zhou, Xiangliang Zhang, and Lichao Sun. 2024b. 1+ 1> 2: Can large language models serve as cross-lingual knowledge aggregators? *arXiv preprint arXiv:2406.14721*.

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. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Anubha Kabra, Sanketh Rangreji, Yash Mathur, Aman Madaan, Emmy Liu, and Graham Neubig. 2023. Program-aided reasoners (better) know what they know. *Preprint*, arXiv:2311.09553.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2024. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, Online. Association for Computational Linguistics.

Geyu Lin, Bin Wang, Zhengyuan Liu, and Nancy F. Chen. 2024. Crossin: An efficient instruction tuning approach for cross-lingual knowledge alignment. *Preprint*, arXiv:2404.11932.

Jiateng Liu, Lin Ai, Zizhou Liu, Payam Karisani, Zheng Hui, Yi Fung, Preslav Nakov, Julia Hirschberg, and Heng Ji. 2025. PropaInsight: Toward deeper understanding of propaganda in terms of techniques, appeals, and intent. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5607–5628, Abu Dhabi, UAE. Association for Computational Linguistics.

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. *Preprint*, arXiv:2310.10378.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. Mapo: Advancing multilingual reasoning through multilingual alignment-as-preference optimization. *Preprint*, arXiv:2401.06838.

Chenkai Sun, Jinning Li, Yi Fung, Hou Chan, Tarek Abdelzaher, ChengXiang Zhai, and Heng Ji. 2023. Decoding the silent majority: Inducing belief augmented social graph with large language model for response forecasting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 43–57, Singapore. Association for Computational Linguistics.

Xiangru Tang, Chunyuan Deng, Hanminwang Hanminwang, Haoran Wang, Yilun Zhao, Wenqi Shi, Yi Fung, Wangchunshu Zhou, Jiannan Cao, Heng Ji, Arman Cohan, and Mark Gerstein. 2024. MIMIR: A customizable agent tuning platform for enhanced scientific applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 486–496, Miami, Florida, USA. Association for Computational Linguistics.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy F. Chen. 2024. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *Preprint*, arXiv:2309.04766.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Shujin Wu, Yi R. Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. 2025a. Aligning LLMs with individual preferences via interaction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7648–7662, Abu Dhabi, UAE. Association for Computational Linguistics.

Shujin Wu, Cheng Qian, Yi R. (May) Fung, Paul Pu Liang, and Heng Ji. 2025b. Plata: Proactive learning with teacher assistance for weak-to-strong generalization.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *Preprint*, arXiv:2306.13063.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024. A survey on multilingual large language models: Corpora, alignment, and bias. *Preprint*, arXiv:2404.00929.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Nianyi Lin, Kaifeng Yun, Linlu Gong, Jianhui Chen, Zhili Wu, Yunjia Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxin Liu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Bin Xu, Jie Tang, and Juanzi Li. 2024. Kola: Carefully benchmarking world knowledge of large language models. *Preprint*, arXiv:2306.09296.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Training and inference configuration

We set $m = 3$ when sampling responses for each of the base models. We finally obtained 17244 and 2168 preference pairs from MEDQA and X-CSQA datasets, respectively. We used LoRA (Hu

Question	The farmer could raise a good chicken, but he wasn't ready to **move up** to a what? A. rooster B. cow C. bought at store D. hen E. tuna *[Ground truth: B. cow]*
Chosen	Explanation: The phrase "move up" is an idiomatic expression that means to progress or advance to the next level or stage. In this context, it implies that the farmer has some experience with raising a chicken (a good one at that), but is **not yet ready to take on more complex or challenging tasks**. Answer: A. rooster *[Incorrect, but the reasoning is more plausible since it realized the answer should be a task more challenging than raising a chicken.]*
Rejected	Explanation: This question tests the development stage of a farmer. The farmer is now able to raise chickens, which shows that he has achieved certain experience and achievements. Well, if he's not ready to upgrade to something, only one of these options makes sense. Answer: D. Hen *[Incorrect, and it does not show the same level of task understanding as the Chosen one]*

Table 4: Qualitative example of CALM generated preference pair, where the chosen answer is not factually correct. The blue text shows the analysis. Although the chosen and rejected samples are both incorrect, the former pays better attention to the key part of the context "**move up**" by mentioning that the farmer will be likely to face a more challenging task. This reasoning shows better context awareness and is more likely to lead to the correct answer.

et al., 2021) Fine-tuning method for DPO and SFT training. The training parameters are listed in Table 5. The inference parameters are shown in Table 6. All the experiments are performed on NVIDIA A100-SXM-80GB GPUs. We utilize the Llama3-8B-Instruct and Mistral-7B-Instruct model from LlamaFactory (Zheng et al., 2024) framework for training and testing.

Parameter	**DPO**	**SFT**
Learning Rate	5e-6	5e-5
num_train_epochs	3.0	3.0
lr_scheduler_type	cosine	consine
per_device_train_batch_size	1	1
warmup_ratio	0.1	0
val_size	0.06	0.06
pref_beta	0.1	-
pref_loss	sigmoid	-
per_device_eval_size	2	2
LoRA_rank	8	8
LoRA_alpha	16	16
LoRA_trainable	q_{proj},v_{proj}	q_{proj},v_{proj}
Optimizer	Adam	Adam

Table 5: DPO, SFT training parameter

Parameter	**Value**
Temperature	1
top_p	0.9
max_new_tokens	512
per_device_eval_batch_size	4

Table 6: Model inference parameters

B Detailed use of the training dataset

B.1 Data source

This section shows the details of the preliminary dataset selection in Section 3.1. 11.6k and 10k multiple choice questions were sampled from the MEDQA-ZH-CN and MEDQA-US question bank (Jin et al., 2020). We also used all the Chinese and English textbooks provided by MEDQA to construct a vector database, which is necessary for the retrieval augmented generation. For X-CSQA (Lin et al., 2021), we sampled 3k Chinese, English, and French questions.

Model	MEDQA(%)			X-CSQA(%)		
	EN	CN	FR	EN	CN	FR
Llama3-8B	58.2	17.1	24.8	52.9	21.5	25.6
Mistral-7B	47.2	18.1	34.7	49.3	21.7	29.0

Table 7: The percentages of positive samples for each language across task settings. English tasks up the largest portion of the positive samples, but there are also considerable amounts of Chinese and French samples.

	EN	**CN**	**FR**
MEDQA	21.4	47.3	31.3
Mistral	20.5	40.0	39.5

Table 8: Percentage of Chinese, French and English language in final CALM training data.

B.2 Statistics of the training datasets

Table 7 and Table 8 shows the percentages of positive samples for each language across task settings. English indeed tasks up the largest portion of the positive samples, but there are still considerable amounts of Chinese and French samples.

B.3 Full result of MEDQA dataset

For MEDQA, we first translate the native Chinese and English questions into other languages, forming a parallel training set in Chinese, English and French. The full testing result of the MEDQA is

Model	MEDQA US				MEDQA CN-ZH			
	Native EN	EN2CN	EN2FR	AVG	Native CN	CN2EN	CN2FR	AVG
Llama3-8B-Instruct	60.1	33.3	44.9	46.1	56.2	59.5	44.7	53.5
+ SFT	62.4 ↑ 2.3	36.1 ↑ 2.8	45.8 ↑ 0.9	47.8 ↑ 1.7	57.1 ↑ 0.9	59.9 ↑ 0.4	46.4 ↑ 1.7	54.5 ↑ 1.0
+ **CALM**	**63.5** ↑ 3.4	**39.8** ↑ 6.5*	**46.3** ↑ 1.4	**49.9** ↑ 3.8	**59.5** ↑ 3.3	**60.8** ↑ 1.3	**47.4** ↑ 2.7	**55.9** ↑ 2.4
Self-RAG	62.6	36.8	46.9	48.8	57.1	59.8	49.1	55.3
+ SFT	63.8 ↑ 1.2	40.3 ↑ 3.5	47.4 ↑ 0.5	50.5 ↑ 0.7	60.3 ↑ 3.2	61.0 ↑ 1.2	51.2 ↑ 3.7	57.5 ↑ 2.2
+ **CALM**	**64.7** ↑ 2.1	**42.6** ↑ 5.8	**49.4** ↑ 2.5	**52.3** ↑ 3.5	**63.7** ↑ 6.6*	**64.3** ↑ 4.5	**52.8** ↑ 3.7	**60.3** ↑ 5.0
Mistral-7B-Instruct	49.8	29.1	38.8	39.2	36.4	47.3	42.4	42.0
+ SFT	50.3 ↑ 0.5	31.6 ↑ 2.5	40.7 ↑ 1.9	40.9 ↑ 1.7	37.9 ↑ 1.5	49.3 ↑ 2.0	44.6 ↑ 2.2	43.9 ↑ 1.9
+CALM	**52.9** ↑ 3.1	**32.7** ↑ 3.6	**41.9** ↑ 3.1	**42.5** ↑ 3.3	**38.5** ↑ 2.1	**51.8** ↑ 4.5	**45.6** ↑ 3.2	**45.3** ↑ 3.3

Table 9: Full result on the translated MEDQA dataset.

Model	EN	FR	ZH-CN	IT	DE	JA
Llama CALM w/ EN	**73.4**	60.8	52.5	61.6	56.5	42.5
Llama CALM w/ EN+FR	**73.6**	**62.0**	52.4	62.0	56.2	43.6
Llama CALM w/ EN+FR+CN	**74.1**	**65.0**	**54.5**	62.3	57.0	44.0

Table 10: We investigate the cross-lingual generalizability by incrementally adding the training languages in CALM and observe the testing result on both trained and untrained languages. Here, in-domain languages (e.g. languages that appeared in the training data) are highlighted in bold font.

illustrated in Table 9. The accuracy is improved across all the languages after CALM tuning, and the native language has the largest performance gain. The performance of non-native languages is possibly constrained by the translation quality.