CALM🧘 : Unleashing the Cross-lingual Self-Aligning Ability of Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are pretrained on extensive multilingual corpora to acquire both language-specific cultural knowledge and universal common sense. Ideally, while LLMs should provide consistent responses to culture-independent questions across languages, we observe significant performance disparities. To address this, we explored 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 language responses and select the most 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.

1 Introduction

LLMs have been pretrained on various knowledge domains in different languages, encoding vast amounts of word knowledge (Yu et al., 2024). This knowledge can be either culture-dependent or culture-independent (Qi et al., 2023). Ideally, LLM should deliver consistent responses to the culture-independent questions. However, due to the imbalance of the pretraining data, this type of knowledge is not well-aligned (Xu et al., 2024). Existing study shows that the model exhibits different proficiency when solving the same task using different languages (Xu et al., 2024; Huang et al., 2024). One possible explanation is that the commonsense knowledge encoded in one language is harder to access when using other languages.

To bridge the gap, recent studies introduced cross-lingual consistency (Qi et al., 2023). In the ideal case, we want different languages to have similar abilities to retrieve internal knowledge so that the model can better generalize in a multilingual setting. Gao et al. (2024) showed that although the multilingual pertaining and instruction tuning contributes to the cross-lingual consistency, it still lacks scalability to improve the cross-lingual knowledge transferability.

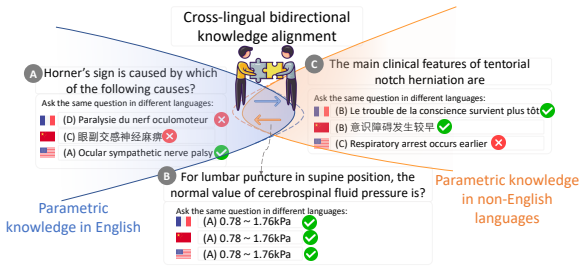

Figure 1: Universal common sense knowledge is often not well-aligned across different languages. (A) It represents knowledge encoded in English that is difficult to retrieve from other languages. (B) is the part of knowledge that is already well-aligned across languages. (C) is the knowledge encoded in other languages that is difficult to retrieve in English.

Chen et al. (2023) utilized translation to construct a multilingual math reasoning instruction dataset. However, it is often labor-intensive to obtain high-quality translation and data annotation. She et al. (2024) leveraged 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 using different languages. Huang et al. (2024) enhanced the multilingual culture commonsense reasoning ability by aggregating the knowledge from diverse languages. In this work, we want to apply this concept of knowledge aggregation to general commonsense reasoning tasks.

To address the challenges of (1) establishing a sustainable method for aligning culture-independent knowledge across different languages

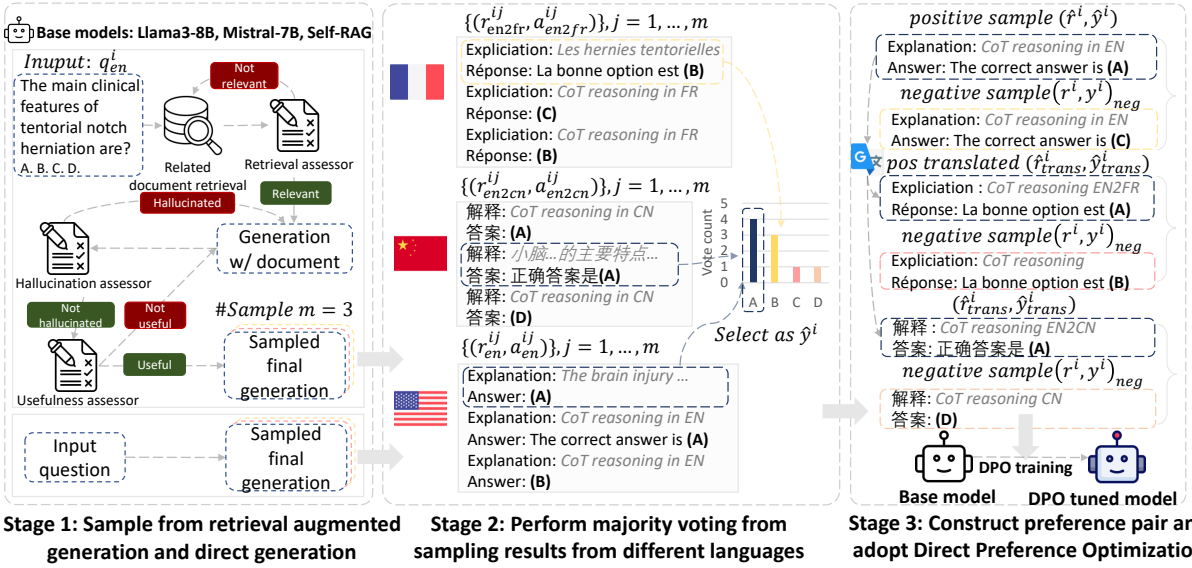

Figure 2: An example of the three stages in our proposed method assuming a question input that is originally in English.

and (2) lacking high-quality annotated data for training, we propose CALM that encourages consistent answers to the same questions in different languages. Especially, this approach utilizes the majority voting results as the alignment target. Note that it does not require a self-consistent answer to be factually correct. As Figure 3 suggested, the self-consistent accuracy (see bar "majority") stably outperforms that of English, making it more suitable to be set as the alignment target. The intuition behind this is suggested by Figure 1. Although English is the most well-performed language, we observed that in our dataset there are still a significant amount of questions answered wrong by English that can be answered right by other languages. This means that only aligning other languages to English may not fully reach the model's full potential of alignment.

Specifically, our approach adopts Direct Preference Optimization (DPO) (Rafailov et al., 2024) for cross-lingual alignment. Our approach consists of three steps. First, we sampled a variety of multilingual CoT outputs from the models. Then, we performed majority voting on the sampled outputs from different languages and selected the answer with the largest number of votes as the positive sample. Finally, we paired the positive sample with all the other answers that were inconsistent with the positive sample, and those pairs were used for DPO training. We also extend this concept to external knowledge sources by integrating Self-supervised Retrieval-Augmented Generation (Self-RAG) (Asai et al., 2023) and DPO techniques. This combined approach allows for the efficient retrieval of the knowledge in the retrieved document.

Experiments were conducted on the challenging MEDQA (Jin et al., 2020) and the X-CSQA (Lin et al., 2021) datasets in Chinese, English, and French. For all the models tested, CALM has yielded an accuracy gain of +3.3%, +4.3%, and +2.6% on MEDQA, and +2.5%, +2.8%, on X-CSQA.

Overall, our key contributions are as follows:

- We proposed a label-free approach to effectively align the knowledge by encouraging cross-lingual consistency, which enables the model to self-improve (Huang et al., 2023).
- We align the culture-independent knowledge by jointly leveraging the knowledge from all the language domains without losing diversity of reasoning inherent to different languages.
- We have tested our approach under zero-shot Chain-of-Thought and retrieval augmented settings on Llama3-8B-Instruct (Dubey et al., 2024), Self-RAG (Asai et al., 2023), and Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), demonstrating the effectiveness of our approach on aligning internal and external knowledge.

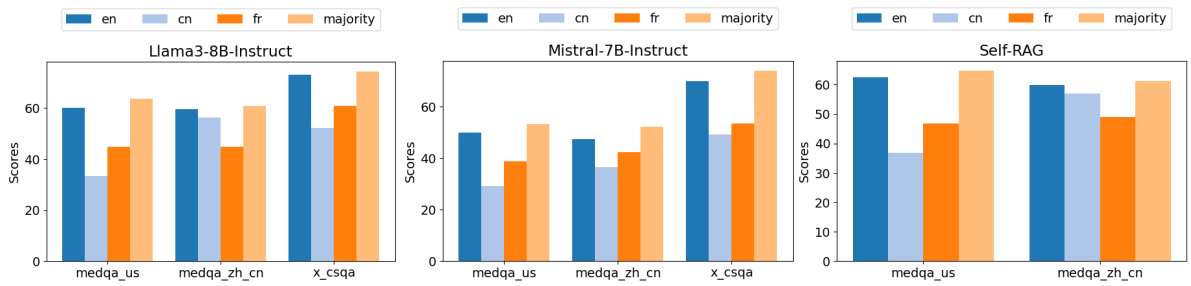

Figure 3: Visualization of single-lingual (EN, CN, FR) accuracy VS the multi-majority voting accuracy. The multilingual majority-voting result always has the highest accuracy.

2 Method

To unleash a greater potential of the model and 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 applied self-consistency (Wang et al., 2022) as the learning signal. By selecting the most voted response as the positive sample with the others, we constructed the preference pairs and adopted DPO to optimize the preference. As the winning response may be from any of the languages, we preserve the diverse knowledge from languages other than English. We also verified our approach in a retrieval-augmented setting, and the result demonstrates that our approach boosts the 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

In this step, we will generate a variety of multilingual reasoning candidates. Given a set of multiple choice questions in native language (e.g. English), denoted as $Q_{en} = \{q^i_{en}\}^{|Q_{en}|}_{i=1}$, we first translate it into two other languages, say Chinese (Q_{en2cn}) and French (Q_{en2fr}). We then apply multiple path decoding with temperature *T = 1* on each variation of the question q^i_{en}, q^i_{en2cn}, q^i_{en2fr} to generate *m* pairs of CoT reasoning paths and answers $\{(r^{ij}_*, y^{ij}_*)\}^m_{j=1}$, wherein y is one of the choice A, B, C or D. The model is prompted to output an "Explanation" following by an "Answer" to better conform with the CoT format (Wei et al., 2024).

2.2 Self-consistency based preference pair construction

The objective of this step is to find pairs of positive (preferred) and negative samples (not preferred) from 2.1 and curate the preference data for DPO training. The most self-consistent answer is identified and will be set as the positive sample. Majority-voting is applied to find the most self-consistent option $\hat{y}_i$ in all the multilingual answers, and it is possible $\hat{y}_i$ is not equal to the ground truth answer.

We can obtain 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 then randomly select one response $(\hat{r}^i, \hat{y}^i)$ from S as the positive sample. Note that this positive sample is not necessarily in English. Hence, we preserve the diverse knowledge of other non-English languages. Negative samples are inconsistent with the positive one, i.e., $y_{negative} \neq \hat{y}^i$. For each positive-negative sample pair, the positive sample will be translated into the language of the negative sample. The final preference pairs from the i-th question were constructed as $p^i = \{p^i_w : (\hat{r}^i{}_{trans}, \hat{y}^i), p^i_l : (r^i, y^i)_{neg}\}$.

2.3 Multilingual knowledge alignment

This step adopts DPO as the alignment approach using the preference pairs (p_w, p_l) obtained from 2.2, where p_w is favored over p_l. For the input question q, the alignment is done by optimizing the following objective:

$$L_{\text{DPO}}(\pi_\theta; \pi_{ref}) = \mathbf{E}(q, p_w, p_l) \sim \mathcal{D}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(p_w|q)}{\pi_{\text{ref}}(p_w|q)} - \beta \log \frac{\pi_\theta(p_l|q)}{\pi_{\text{ref}}(p_l|q)}\right)\right] \quad (1)$$

Model	MEDQA			X-CSQA			
	EN	ZH-CN	AVG	EN	ZH-CN	FR	AVG
Llama3-8B-Instruct	60.1	56.2	58.2	73.1	52.1	60.8	62.0
+ SFT	62.4	57.1	59.8	73.8	53.2	62.3	63.1
+ CALM (Ours)	**63.5**	**59.5**	**61.5**	**74.1**	**54.5**	**65.0**	**64.5**
Self-RAG	62.6	57.1	59.9	-	-	-	-
+ SFT	63.8	60.3	62.1	-	-	-	-
+ CALM (Ours)	**64.7**	**63.7**	**64.2**	-	-	-	-
Mistral-7B-Instruct	49.8	36.4	43.1	70.0	49.1	53.5	57.5
+ SFT	50.3	37.9	44.1	71.4	50.9	54.7	59.0
+ CALM (Ours)	**52.9**	**38.5**	**45.7**	**73.5**	**52.2**	**55.1**	**60.3**

Table 1: Model accuracy on the test set of MEDQA and X-CSQA, respectively. "AVG" denotes the average accuracy of English, Chinese, and French, which indicates the overall domain knowledge of the model. The accuracy gain in each non-AVG column can be interpreted as a knowledge gain in a specific language. The bold text represents the best result in the given model. Note there is no X-CSQA results for Self-RAG because there are no documents available for retrieval.

3 Experiment and Results

3.1 Task definition

We demonstrated the effectiveness of our method on three types of tasks: (1) Direct reasoning on college-level medical questions, (2) Noisy evidence retrieval (Jin et al., 2020) and complex reasoning over multiple evidence on college-level medical questions, and (3) General commonsense reasoning.

3.2 Baselines

Our experiments trained three base models, which are Llama3-8B-Instruct, Mistral-7B-Instruct-v0.2, and Self-RAG. The testing results of the former two models demonstrated the effectiveness of our approach in terms of internal knowledge alignment and the latter one in terms of external knowledge. The primary baseline is the direct inference results sampled from all the base models. To prove the necessity of DPO in this alignment task, we further adopted supervised fine-tuning (SFT) on all the positive samples as our second baseline. It is worth mentioning that we adopted continued learning during SFT, meaning that different languages are iteratively trained one by one. This type of training is more efficient in a way that models can be continuously exposed to knowledge inconsistency.

3.3 Results

We use the accuracy score of the multiple choice question as the evaluation metric, which is an indicator of the model's knowledge in different languages. As is shown in Table 1, our DPO method has outperformed the base model and the supervised fine-tuned model under all settings. On average, CALM consistently outperformed the base model for more than 2.5%, confirming its effectiveness. We also observed that the accuracy of English, which is the best-performing language, has also been increased by a similar level as other languages, meaning that English can also learn from other languages. In all the experiment settings, CALM has yielded an accuracy gain of +3.3% on Llama3-8B-Instruct, +4.3% on Self-RAG, and +2.6% on Mistral-7B-Instruct-v0.2 for the experiments on MEDQA dataset. Also, CALM Llama3-8B-Instruct and CALM Mistral-7b-Instruct-v0.2 has also experienced a boost of +2.5% and +2.8% on X-CSQA dataset compared with the baseline. Therefore, we can conclude that our approach has successfully facilitated the cross-lingual self-alignment.

4 Conclusion

In this paper, we introduce CALM, a novel framework designed to facilitate the alignment of knowledge across different languages within LLMs. Through comprehensive experiments, we demonstrate the effectiveness of CALM in achieving robust cross-lingual knowledge alignment. Leveraging the LLM's own predictions as a reward signal, we investigate its capacity for self-alignment and self-improvement. Notably, we even observe knowledge transfer from lower- to higher-performing languages. These results highlight the effectiveness of our method in enhancing cross-lingual alignment and reveal the multilingual self-alignment potential of LLMs.

Limitations

One of the main limitations of our study is that due to the limitation of the computational resources, we are unable to carry out experiments on larger models. For the same reason, we are also not able to perform full-parameter fine-tuning on the models and can only use LoRA DPO fine-tuning as an alternative. The translations in the experiment were all done by Google Translate API, which may not be accurate sometimes because the dataset contains a lot of challenging medical terminology, which may have hindered our final performance. For the DPO training data construction, since the majority-voting accuracy only surpasses the accuracy of monolingual (e.g. English) by a small amount, the final alignment performance may be constrained by the noisy labels in the positive samples. Future work can further improve the majority voting accuracy.

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, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein,

Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *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.

Yue Huang, Chenrui Fan, Yuan Li, Siyuan Wu, Tianyi Zhou, Xiangliang Zhang, and Lichao Sun. 2024. 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 do-

main question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

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.

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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv 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.

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.

A DPO training 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 et al., 2021) Fine-tuning method for DPO training. The learning rate was set to be 6e-05. The number of training epochs was 3. A LoRA rank of 8 and a LoRA alpha of 16 were used for the LoRA DPO fine-tuning.

B Detailed use of the training dataset

This section shows the details of the priliminary dataset selection in Section 3.1. For the task (1) and (2) of 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 task (3), we sampled 3k Chinese, English, and French questions in the X-CSQA dataset (Lin et al., 2021).