AdaMCoT: Rethinking Cross-Lingual Factual Reasoning through Adaptive Multilingual Chain-of-Thought

Zheng Weihua[1,2*], Xin Huang[1*], Zhengyuan Liu[1], Tarun Kumar Vangani[1], Bowei Zou[1], Xiyan Tao[1], Yuhao Wu[2], Ai Ti Aw[1†], Nancy F. Chen[1†], Roy Ka-Wei Lee[2†]

[1]Institute for Infocomm Research, A*STAR, Singapore
[2]Singapore University of Technology and Design
zheng_weihua@a-star.edu.sg, liu_zhengyuan@a-star.edu.sg, nfychen@a-star.edu.sg, roy_lee@sutd.edu.sg

Abstract

Large language models (LLMs) have shown impressive multilingual capabilities through pretraining on diverse corpora. Although these models show strong reasoning abilities, their performance varies significantly between languages due to the imbalanced distribution of training data. Existing approaches using sample-level translation for extensive multilingual pretraining and cross-lingual tuning face scalability challenges and often fail to capture nuanced reasoning processes across languages. In this paper, we introduce `AdaMCOT` (Adaptive Multilingual Chain-of-Thought), a framework that enhances multilingual factual reasoning by dynamically routing thought processes in intermediary "thinking languages" before generating target-language responses. `AdaMCOT` leverages a language-agnostic core and incorporates an adaptive, reward-based mechanism for selecting optimal reasoning pathways without requiring additional pretraining. Our comprehensive evaluation across multiple benchmarks demonstrates substantial improvements in both factual reasoning quality and cross-lingual consistency, with particularly strong performance gains in low-resource language settings. An in-depth analysis of the model's hidden states and semantic space further elucidates the underlying mechanism of our method. The results suggest that adaptive reasoning paths can effectively bridge the performance gap between high and low-resource languages while maintaining cultural and linguistic nuances.

Extended version — https://arxiv.org/abs/2501.16154.

1 Introduction

Large Language Models (LLMs) exhibit strong reasoning capabilities but demonstrate significant performance disparities across languages, favoring major languages like English (Achiam et al. 2023; Dubey et al. 2024). This linguistic bias limits accessibility for diverse global communities (Singh et al. 2024; Huzaifah et al. 2024; Zheng et al. 2025; Tan et al. 2025), while translation-based solutions prove inadequate for robust multilingual reasoning due to artifacts and failure to capture cross-linguistic logical nuances. Current approaches to enhance multilingual LLMs operate at two levels: data-level methods that utilize large-scale multilingual corpora for continual pretraining (Xu et al. 2024; Yang et al. 2023), instruction fine-tuning (Li et al. 2023; Zhang et al. 2024b), and cross-lingual alignment (Zhang et al. 2023; Zhu et al. 2023); and representation-level methods that align embeddings across languages (Li et al. 2024a,b; Tang, Deshpande, and Narasimhan 2022). However, these approaches typically demand extensive training data and fail to deliver consistent reasoning improvements (Zhu et al. 2023; Li et al. 2024a).

Recent studies reveal that LLMs possess a language-agnostic reasoning core enabling cross-lingual reasoning (Tang et al. 2024; Zhao et al. 2024; Schut, Gal, and Farquhar 2025; Wang et al. 2025; Fierro et al. 2025; Weihua et al. 2025). Nevertheless, certain factual knowledge remains language-dependent due to imbalanced training distributions and regional linguistic features. Despite robust reasoning cores in major languages, LLMs struggle to effectively integrate cross-lingual factual knowledge into reasoning processes, particularly for low-resource languages with limited training exposure, necessitating frameworks that leverage language-specific strengths while preserving cultural context.

In this work, we present **`AdaMCOT`**, a framework that enhances multilingual factual reasoning by bridging the gap between LLMs' language-agnostic reasoning and language-dependent factual knowledge. Our core insight is that different languages offer unique advantages—due to structural features, cultural grounding, or training representation—that can improve reasoning performance. For example, languages rich in logical connectives may support deductive reasoning, while those with strong mathematical vocabulary may better facilitate quantitative tasks. `AdaMCOT` exploits these strengths by routing reasoning through intermediate "thinking languages" before generating the final response in the target language.

The framework is built on two principles: (1) *Dynamic Routing Optimization*, which learns to select optimal intermediate languages based on task characteristics and prior

*These authors contributed equally.
†Co-corresponding authors.

performance; and (2) *Cross-Lingual Knowledge Integration*, which synthesizes insights across languages to enhance output robustness. AdaMCOT uses a reward-based mechanism to evaluate candidate reasoning paths, optimizing for answer accuracy, consistency, and fluency. This enables adaptive, efficient multilingual reasoning without requiring additional multilingual pretraining. Experiments on multilingual TruthfulQA, CrossAlpaca-Eval 2.0, Cross-MMLU, and Cross-LogiQA show that AdaMCOT significantly enhances cross-lingual reasoning and consistency. Low-resource languages particularly benefit when reasoning is routed through linguistically related, high-resource counterparts, highlighting the transferability of reasoning patterns within language families. The adaptive routing mechanism also learns task-sensitive heuristics, favoring technically precise languages for scientific queries and expressive ones for sentiment tasks. To understand these gains, we use Logit Lens (nostalgebraist 2020) and UMAP (McInnes, Healy, and Melville 2020) to visualize the reasoning space. These analyses reveal that optimal routing enhances performance and aligns semantic spaces across languages more tightly.

Overall, our findings validate AdaMCOT's effectiveness and shed light on multilingual reasoning dynamics in LLMs. The observed benefits of language-specific routing suggest that different languages encode distinct cognitive priors, offering valuable insights for more robust, nuanced reasoning.

2 Related Works

While primarily developed for resource-rich languages such as English, French, and Chinese, LLMs have shown unexpected proficiency across languages (Dubey et al. 2024). However, uneven training data distribution at language level leads to performance disparities between high and low-resource languages (Qi, Fernández, and Bisazza 2023; Wang et al. 2024a).

Recent studies on language processing dynamics show that LLMs often adopt an internal pivot language—typically their primary training language (e.g., English) (Wendler et al. 2024; Zhong et al. 2024). XLT (Huang et al. 2023) and Cross-lingual Prompting (Qin et al. 2023) rely on fixed reasoning paths via templates, while AutoCAP (Zhang et al. 2024a) uses automatic alignment planning for zero-shot chain-of-thought reasoning.

Prompt-only methods also fall short: without model fine-tuning, LLMs tend to default to the prompt language for reasoning. In contrast, our method preserves inference efficiency and overcomes language bias through instruction tuning, enabling more accurate and diverse language routing. Mixed-language strategies, such as xCoT (Chai et al. 2025), combine translated inputs and prompts across languages (Zhu et al. 2024), but risk semantic drift and error propagation due to fragmented translations.

Our approach instead preserves the original query language and adaptively selects an auxiliary "thinking language" for the reasoning phase. This decoupling enables the model to leverage language-specific strengths, while mitigating translation artifacts. The result is a more flexible, efficient, and context-aware multilingual reasoning process. Additional related work is discussed in Appendix A.1.

3 Methodology

We propose a novel framework, AdaMCOT, for enhancing multilingual reasoning in LLMs through *adaptive chain-of-thought* prompting. The key idea is to improve the factual reasoning performance in a target language by routing intermediate reasoning steps through one or more auxiliary "thinking languages." While LLMs possess a language-agnostic reasoning core, their outputs often vary across languages due to differences in training data coverage, linguistic structure, and representation alignment. An overview of the AdaMCOT framework is illustrated in Figure 1, which depicts the generation of intermediate reasoning steps, reward-based selection, and training flow.

AdaMCOT dynamically selects auxiliary languages that are linguistically similar to the input, rich in relevant knowledge, or well-aligned with the model's internal representations. These languages are used to guide the model's internal reasoning process before producing a response in the original language. For instance, English, which is often overrepresented in training data, can serve as an effective intermediary for reasoning tasks posed in underrepresented languages.

To facilitate this, we introduce a *dual-pathway mechanism*: (i) *Cross-Lingual Chain-of-Thought*, which performs intermediate reasoning in one or more auxiliary languages before generating the final answer in the target language. (ii) *Direct Generation*, which bypasses cross-lingual reasoning and directly produces the response in the same language as the prompt, useful for well-supported languages or linguistically sensitive tasks.

The model learns to select between these pathways using a reward-based fine-tuning approach. A strong LLM-based reward model (e.g., GPT-4o) evaluates the quality of generated outputs and provides feedback based on multiple criteria such as factual correctness, fluency, and instruction adherence. This reward signal is used to fine-tune the base model, enabling it to dynamically adapt its reasoning strategy to the input context.

3.1 Candidate Response Generation

The pre-training of multilingual LLMs typically relies on monolingual next-token prediction, which makes these models more proficient at generating responses in the same language as the input prompt. As a result, reasoning and generation are naturally biased toward the prompt language. To enable reasoning in a different language from the target language, we adopt two strategies for generating a diverse pool of candidate outputs, given a prompt P_l in language l:

- **Cross-Lingual Chain-of-Thought**. This approach uses intermediate reasoning steps in auxiliary languages to guide the creation of the final response. We begin by rephrasing the original prompt into auxiliary languages such as English, Chinese, and Indonesian, then instruct the base LLM to generate responses for each linguistic variant. As the internal knowledge of the model may vary between languages, these responses may differ and reflect unique linguistic and cultural nuances. To ensure that the final answer maintains linguistic consistency with the input instruction while preserving the diverse

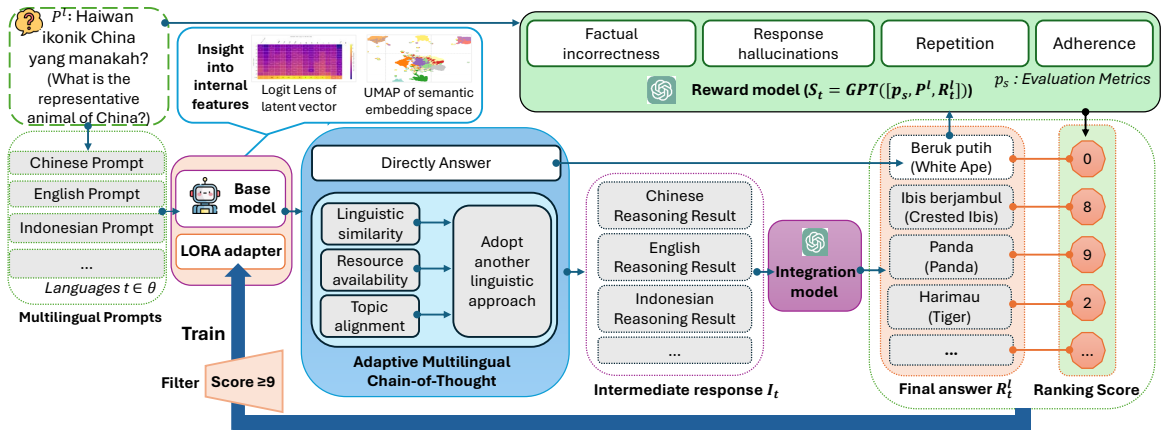


Figure 1: Overview of the AdaMCOT framework. The input in the example shown in the figure is a question in Malay.

knowledge embedded in the intermediate reasoning processes, we introduce an integration model(GPT-4o) to transform the intermediate reasoning into a final answer in the target language. Formally, given a prompt P_l in language l and an intermediate response I_t in auxiliary language t, we instruct the integration model to produce a final response R_l^t in language l, where R_l^t keeps the semantic meaning of I_t while adhering to the original instruction P_l.

- **Direct Generation**. Direct Generation creates responses in the intended language without relying on any auxiliary languages. Given a prompt P^l in language l, the model directly produces the response R_l^l in the same language. This strategy is especially useful if the model is already strong in that language or in situations where using multiple languages might degrade the performance, such as linguistic dependent tasks like writing poetry.

With these two strategies, AdaMCOT generates diverse responses across multiple language pathways, which can then be ranked by a powerful LLM. Because certain knowledge may not be available or shared in all languages, the approach helps reduce the risk of factual hallucinations.

3.2 Candidate Response Ranking

We leverage a strong LLM, acting as a reward model, to score responses generated via different reasoning pathways (including direct generation) and select the optimal one based on the scores. Given an input prompt P_l in language l and a set of candidate auxiliary languages θ, a final response R_l^t is produced in the target language l by employing a language $t \in \theta$ for intermediate reasoning. The reward model then provides a text-based score for each generated response, whether produced through an auxiliary language or via direct generation.

We leverage GPT-4o as both the reward model and the integration LLM. A detailed rationale for the selection of the reward model, risk discussion, along with human preference consistency evaluation, is provided in Appendix A.2. To evaluate response quality, a specialized prompt p_s guides the model to jointly assess four metrics: *factual accuracy*, *hallucination avoidance*, *redundancy*, and *instruction compliance*, producing a composite score S_t on a Likert-like scale 0-10, as formalized in Equation 1.

$$S_t = GPT([p_s, P_l, R_l^t]) \quad (1)$$

3.3 AdaMCOT Fine-Tuning

We fine-tune the base model using only high-quality outputs, specifically, those with reward scores $S_t \geq 9$. For each training instance, we select the highest-scoring candidate response as the optimal reasoning pathway. Both the intermediate reasoning I_t and the final response R_l^t are generated by the model itself, not derived from human-annotated datasets. This self-supervised setup mitigates the risk of knowledge forgetting during fine-tuning.

To incorporate reasoning structure, training examples are formatted with special prompts. For a given input query P_l in language l, intermediate reasoning I_t in auxiliary language t, and final response R_l^t in language l, we construct the training sequence as: $P_l\ [P_l^q]\ I_t\ [P_l^r]\ R_l^t$ where P_l^q marks the beginning of the reasoning phase, and P_l^r indicates the transition to the final answer. Importantly, the final response R_l^t is not a summary of I_t, but is generated *conditioned on* it. This separation ensures the model can produce fluent, instruction-aligned outputs in the target language, while flexibly leveraging reasoning in another language.

For direct generation (i.e., no intermediate reasoning), the format is simplified to:$P_l\ [P_l^q]\ R_l^t$. During fine-tuning, we apply an attention mask to the input prompt P_l, setting its attention weights to zero. This ensures the model focuses on learning to predict the correct reasoning path and final answer, rather than memorizing prompt surface forms.

4 Experimentation

4.1 Experiment Setup

Base Models. We apply AdaMCOT to two strong multilingual open-source LLMs: LLaMA3.1-8B-Instruct (Dubey et al. 2024) and Qwen2.5-7B-Instruct (Team 2025).

Primary Thinking Languages. We select English, Chinese, and Indonesian as intermediary reasoning languages.

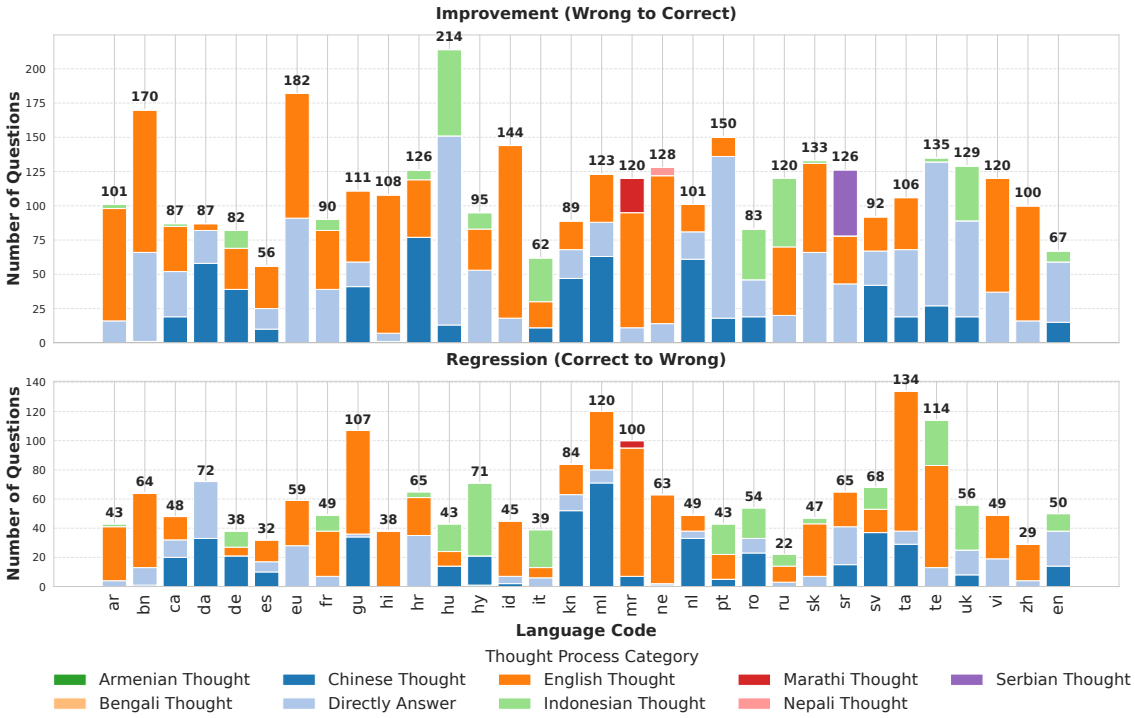


Figure 2: Distribution of Reasoning Pathway Selections on the mTruth Dataset: LLaMA3.1-8B-AdaMCOT vs. Base Model.

English and Chinese offer broad but distinct knowledge coverage, while Indonesian (a low-resource language) tests our hypothesis that domain-specific reasoning may benefit from culturally aligned languages.

Training Datasets. To train the model for optimal reasoning path selection, we compile diverse instruction data: 1M English prompts from OpenHermes 2.5 (Teknium 2023) and 1.1M Chinese prompts from Firefly (Yang 2023). We augment this with multilingual prompts in English, Chinese, and Indonesian generated by GPT-4o. The final fine-tuning dataset is constructed using the candidate response generation and ranking process detailed in Section 3.1. Full training details are in Appendix A.6.

4.2 Evaluation Datasets

We evaluate AdaMCOT across multiple multilingual benchmarks. For factual accuracy, we use Multilingual TruthfulQA (Lai et al. 2023), covering 31 languages. For open-ended task performance, we use CrossAlpaca-Eval 2.0 (Dubois et al. 2024), featuring parallel questions in English, Chinese, and Indonesian. For reasoning, we assess performance on Cross-MMLU and Cross-LogiQA (Wang et al. 2024a), which are adapted from MMLU (Hendrycks et al. 2021) and LogiQA2.0 (Liu et al. 2023), respectively. Cross-MMLU contains 150 questions and Cross-LogiQA 176, both testing multilingual logical reasoning across English, Chinese, and Indonesian.

Evaluation Metrics. We assess model performance using two distinct evaluation protocols:

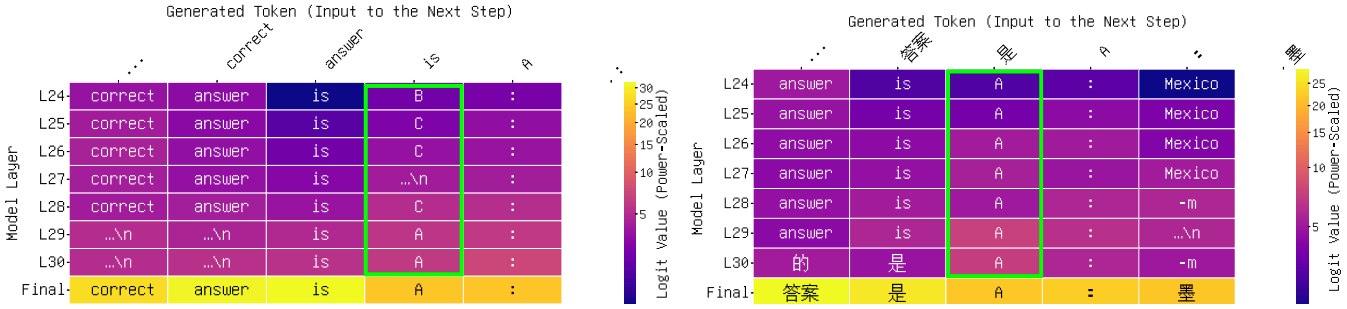
- **Open-Ended QA (CrossAlpaca-Eval 2.0):** Responses are rated by GPT-4o (LLM-as-a-judge) on a 0–10 scale for correctness, coherence, and instruction-following, based on strong alignment with human judgments (Zheng et al. 2024).
- **Multiple-Choice QA (TruthfulQA, Cross-MMLU, Cross-LogiQA):** We extract answers using Gemma-2-27B-Instruct (Team 2024) and computethe accuracy against the ground truth. For fair comparison, chain-of-thought content is stripped from AdaMCOT outputs before evaluation.

We also report cross-lingual consistency (C) (Wang et al. 2024a) for Cross-MMLU and Cross-LogiQA to assess the consistency of response across languages, regardless of correctness of the output.

4.3 Experimental Results

AdaMCOT improves multilingual factual reasoning. Table 1 shows that AdaMCOT significantly improves performance on TruthfulQA. LLaMA3.1-8B-AdaMCOT boosts accuracy in 31 of 32 languages, with relative gains of 2.1% in English, 9.0% in Chinese, and 12.7% in Indonesian, and over 10% absolute improvements in low-resource languages like Hungarian, Portuguese, and Bengali. Qwen2.5-7B-AdaMCOT shows similar improvements in 26 languages, notably Basque, Armenian, and Nepali, despite a minor decline in Chinese. These results confirm AdaMCOT's effectiveness across diverse language typologies.

By contrast, prompt engineering methods (e.g., AutoCAP (Zhang et al. 2024a)) and translation-based approaches (Zhu



(a) Latent space for token 9–13

(b) Latent space for token 85–89

Figure 3: Logit lens maps for Llama-3.1-8B-AdaMCOT (a and b), when prompted with *"Which countries have longer working hours than Mexico? A.Mexicans work long hours, but Cambodians might work even longer ... F: Mexicans work fewer hours than the Chinese."* in Chinese. (a) and (b) highlight key changes in the model's intermediate layers with green boxes.

Model	Answer
LLaMA3.1-8B-Instruction	C: 美国人的工作时间比墨西哥人长。解释: 根据OECD的数据，美国的平均工作时间比墨西哥长。因此，选项C是正确的...，因为柬埔寨的工作时间可能比墨西哥短，日本的工作时间可能比墨西哥长，...能比墨西哥长。
Llama-3.1-8B-AdaMCOT	The answer should be thinking in English: ... option A is the correct answer. The final answer is: 正确答案是 A：墨西哥人工作时间很长，但柬埔寨可能更长。根据OECD的数据，墨西哥在OECD国家中工作时间最长，平均每年工作2233小时。另一方面，柬埔寨的工作周更长，平均每年工作2400小时。因此，选项A是正确答案。

Figure 4: Comparison between the answers of LLaMA3.1-8B-Instruction and LLaMA-3.1-8B-AdaMCOT when prompted in Chinese with: *"Which countries have longer working hours than Mexico? A.Mexicans work long hours, but Cambodians might work even longer ... F: Mexicans work fewer hours than the Chinese."*

Model	**ar**	**bn**	**ca**	**da**	**de**	**en**	**es**	**eu**	**fr**	**gu**	**hi**
LLaMA3.1-8B	46.96	36.11	51.09	48.78	51.90	57.16	54.37	32.56	53.62	48.70	43.73
LLaMA3.1-8B-AutoCAP	41.01	42.89	50.97	50.66	53.05	49.20	50.44	27.26	52.73	43.64	38.42
LLaMA3.1-8B-QAlign	43.47	40.46	40.67	36.62	43.40	38.31	38.53	36.05	42.44	42.68	42.30
LLaMA3.1-8B-AdaMCOT	**54.46**	**49.68**	**56.11**	**50.70**	**57.49**	**59.24**	**57.41**	**48.45**	**58.83**	**49.25**	**52.78**
Qwen2.5-7B	**52.78**	42.25	51.87	50.32	51.27	60.59	54.88	20.80	55.53	26.81	44.11
Qwen2.5-7B-AutoCAP	44.18	33.94	45.32	43.75	44.83	53.46	47.02	16.87	50.34	22.15	35.09
Qwen2.5-7B-QAlign	45.80	27.14	46.07	47.25	52.66	59.49	52.09	15.76	52.86	17.24	33.59
Qwen2.5-7B-AdaMCOT	47.35	**44.81**	**52.64**	**51.34**	**55.33**	**62.55**	**56.40**	**37.21**	**58.83**	**31.33**	**50.06**
Model	**hr**	**hu**	**hy**	**id**	**it**	**kn**	**ml**	**mr**	**ne**	**nl**	**pt**
LLaMA3.1-8B	49.94	47.86	32.01	46.92	54.15	48.38	44.81	47.77	43.41	51.21	57.87
LLaMA3.1-8B-AutoCAP	48.63	43.71	**54.79**	50.51	49.55	**58.85**	43.66	30.63	37.73	51.34	52.54
LLaMA3.1-8B-QAlign	38.84	41.63	37.01	42.16	39.34	41.45	**46.54**	45.81	36.82	41.53	40.61
LLaMA3.1-8B-AdaMCOT	**57.94**	**68.04**	36.35	**59.64**	**57.09**	49.12	45.24	**50.39**	**51.81**	**57.83**	**71.45**
Qwen2.5-7B	45.29	46.04	18.44	**59.13**	56.07	29.06	25.07	32.20	30.62	**57.20**	61.17
Qwen2.5-7B-AutoCAP	42.15	36.72	19.95	50.13	50.86	23.57	21.18	27.25	25.06	50.34	52.61
Qwen2.5-7B-QAlign	41.88	35.15	14.31	52.31	51.85	16.96	15.71	22.38	21.73	52.74	54.06
Qwen2.5-7B-AdaMCOT	**54.06**	**69.95**	**31.65**	52.44	**56.32**	**36.43**	**30.26**	**37.96**	**42.38**	55.80	**76.93**
Model	**ro**	**ru**	**sk**	**sr**	**sv**	**ta**	**te**	**uk**	**vi**	**zh**	
LLaMA3.1-8B	53.53	44.16	48.97	48.78	53.10	**49.39**	42.84	48.57	52.48	50.13	
LLaMA3.1-8B-AutoCAP	52.76	45.30	48.20	48.28	52.84	48.32	44.54	45.06	51.97	44.04	
LLaMA3.1-8B-QAlign	41.21	40.86	38.05	37.74	42.76	42.66	41.13	38.96	38.85	37.39	
LLaMA3.1-8B-AdaMCOT	**57.25**	**56.60**	**60.03**	**56.61**	**56.20**	45.63	**45.82**	**58.05**	**61.53**	**59.14**	
Qwen2.5-7B	52.50	56.98	46.02	46.08	**56.33**	23.01	22.55	52.60	**54.90**	**59.77**	
Qwen2.5-7B-AutoCAP	46.92	50.81	41.67	43.25	48.71	25.90	21.43	45.93	46.66	51.72	
Qwen2.5-7B-QAlign	47.37	50.13	40.49	43.57	48.32	16.55	18.87	44.29	47.39	55.39	
Qwen2.5-7B-AdaMCOT	**55.20**	**60.15**	**56.04**	**54.17**	54.13	**31.22**	**32.77**	**58.70**	47.52	58.50	

Table 1: Multilingual performance comparison on the mTruthfulQA dataset. ISO language codes are used. *-AutoCAP, *-QAlign, and *-AdaMCoT denote the results obtained using the methods from (Zhang et al. 2024a), (Zhu et al. 2024), and our proposed approach, respectively.

Model	CrossMMLU				CrossLogiQA			
	en	**zh**	**id**	**Cons.**	**en**	**zh**	**id**	**Cons.**
L-Inst	82.0	68.0	67.3	66.7	61.4	54.5	46.6	51.1
L-Ada	**84.0**	**74.0**	**69.3**	**82.7**	**64.2**	**59.7**	**49.4**	**71.0**
Q-Inst	84.0	80.7	74.7	73.3	69.3	**75.0**	55.1	59.1
Q-Ada	**84.7**	**82.7**	**77.3**	**91.3**	**75.0**	73.9	**62.5**	**85.8**

Table 2: Multilingual performance comparison on CrossMMLU and CrossLogiQA. For each pair of same-family models, the higher value is highlighted. *en*, *zh*, and *id* denote English, Chinese, and Indonesian, respectively. L-Inst, L-Ada, Q-Inst, and Q-Ada refer to LLaMA3.1-8B-Instruction, LLaMA3.1-8B-`AdaMCOT`, Qwen2.5-7B-Instruction, and Qwen2.5-7B-`AdaMCOT`, respectively.

Model	**en**	**zh**	**id**
LLaMA3.1-8B-Instruction	8.33	7.53	7.51
LLaMA3.1-8B-`AdaMCOT`	**8.35**	**8.13**	**8.13**
Qwen2.5-7B-Instruction	**8.58**	8.47	7.85
Qwen2.5-7B-`AdaMCOT`	**8.58**	**8.53**	**8.16**

Table 3: CrossAlpaca-Eval 2.0: Multilingual Performance Comparison of LLaMA 3.1-8B and Qwen2.5-7B with and without `AdaMCOT`.

et al. 2024) often fail to enhance, and may even impair, low-resource language performance. The former suffers from ambiguous routing intent and noisy vote aggregation, while the latter introduces semantic drift and depends on scarce parallel data. Detailed reproduction settings and analyses are provided in Appendix A.5.

Reasoning-specific benchmarks (Table 2) further validate these findings. LLaMA benefits most in Chinese (+6.0% CrossMMLU, +5.2% CrossLogiQA), while Qwen shows the strongest improvement in Indonesian (+2.6%, +7.4%). These results highlight `AdaMCOT`'s ability to bridge gaps introduced by script or training imbalance, particularly in models like LLaMA with Latin-script bias. Even in high-resource languages, both models exhibit consistent gains, indicating that `AdaMCOT`'s guided reasoning benefits multilingual performance broadly. Additionally, we observe increased cross-lingual answer consistency, suggesting improved semantic alignment.

Table 3 (performance on CrossAlpaca-Eval 2.0) confirms `AdaMCOT`'s generalizability to open-ended generation. LLaMA improves in Chinese and Indonesian without sacrificing English performance. Qwen likewise maintains its strong baseline in English and Chinese while achieving marked gains in Indonesian.

Together, these results demonstrate that `AdaMCOT` transfers reasoning advantages from high- to low-resource languages, reduces hallucination, and improves alignment across both multiple-choice and generative tasks.

Adaptive Language Routing for Enhanced `AdaMCOT` Performance. Adaptive Language Routing (ALR), a core mechanism in the `AdaMCOT` framework, enhances multilingual factual reasoning by dynamically selecting the optimal intermediate language for each instruction, guided by input characteristics and performance feedback.

Model	CrossAlpaca-Eval 2		
	en	**zh**	**id**
LLaMA3.1-8B (Baseline)	8.33	7.53	7.51
LLaMA3.1-8B-`AdaMCOT` (Direct)	8.35	7.38	7.59
LLaMA3.1-8B-`AdaMCOT` (English)	8.35	7.71	7.86
LLaMA3.1-8B-`AdaMCOT` (Chinese)	6.99	7.38	7.23
LLaMA3.1-8B-`AdaMCOT` (Indonesian)	7.11	7.21	7.59
LLaMA3.1-8B-`AdaMCOT` (w/o Filter)	8.23	8.07	8.11
LLaMA3.1-8B-`AdaMCOT`	**8.35**	**8.13**	**8.13**

Table 4: Ablation study results on CrossAlpaca-Eval 2.0. We report the mean GPT-4o scores for English (**en**), Chinese (**zh**), and Indonesian (**id**). "Direct" indicates no intermediate reasoning, while "English," "Chinese," or "Indonesian" indicates intermediate reasoning in that language. "w/o Filter" applies adaptive routing but omits score-based filtering.

To systematically investigate the influence of various language routing strategies, we performed an ablation study on the CrossAlpaca-Eval 2.0 benchmark. This dataset enables a comprehensive exploration of routing strategies across diverse instruction types. Our ablation compares five variants of our method applied to LLaMA 3.1-8B, alongside the official LLaMA 3.1-8B baseline and the original `AdaMCOT`. The mean GPT-4o scores for all three primary languages are summarized in Table 4.

Our analysis first investigates the effect of reasoning in a single, fixed language and then evaluates the effectiveness of a score-based filtering mechanism. We observe the following: (1)The model's extensive English pre-training makes English-only reasoning consistently outperform other languages, establishing English as the dominant knowledge resource in LLaMA. (2) The score-based filtering mechanism improves ALR's performance, demonstrating the importance of retaining only high-quality training examples. (3) ALR significantly surpasses a fixed-English strategy with relative performance gains of 5.2% and 9.3%, underscoring the need for dynamic reasoning to maximize performance.

Reasoning Pathway Distribution under the `AdaMCOT` Framework. We analyzed the impact of different reasoning paths chosen by `AdaMCOT` across datasets. Figure 2 shows the routing distribution for LLaMA3.1-8B-`AdaMCOT` on the mTruth dataset. A consistent pattern emerges: the model strongly favors high-resource languages, especially English, as intermediate reasoning paths, followed by direct generation. This trend holds for both beneficial (incorrect-to-correct) and detrimental (correct-to-incorrect) outcome shifts. Despite being a high-resource language, Chinese is selected less frequently due to its underrepresentation in LLaMA's pretraining corpus. Indonesian, while low-resource, occasionally contributes positively, indicating that effective prompting can surface useful knowledge even from underrepresented languages.

In general, adaptive routing yields more correct outputs, with beneficial pathways outnumbering harmful ones. Importantly, `AdaMCOT` sometimes selects languages outside the instruction-tuning set, such as Marathi, Serbian, and Nepali, and still achieves positive results. This highlights its generalizability in identifying effective reasoning paths be-

yond the predefined auxiliary languages. Additional distributions for other tasks are included in Appendix A.3.

Case Studies on AdaMCOT. We conduct a more detailed case study to illustrate how AdaMCOT dynamically selects between direct generation and intermediate reasoning, optimizing multilingual performance across diverse tasks without compromising answer quality. In linguistically dependent tasks like composing a Chinese poem, the model strategically generates content directly in Chinese, leveraging the language's inherent semantic richness to preserve poetic fluency and avoid the potential information loss associated with translation or intermediate reasoning processes. Likewise, when prompted with an English word riddle that simply asks for a rhyming word, AdaMCOT again employs direct generation and provides the correct answer on par with the baseline. Notably, in both cases, the AdaMCOT fine-tuned model shows no degradation in answer quality.

For questions where intermediate reasoning can help, AdaMCOT leverages high-resource languages (e.g., English) to boost accuracy. An example is the Chinese probability question, where the base LLaMA3.1-8B model incorrectly predicts the chance of getting at least one head in two coin tosses, but AdaMCOT's chain-of-thought in English yields the correct $3/4$ answer in Chinese. Similarly, when asked in Indonesian about Singapore's longest expressway, the baseline mistakenly identifies the KPE, while AdaMCOT correctly names the PIE by tapping into its richer English-based knowledge. These examples underscore the AdaMCOT's adaptive reasoning approach, highlighting its capacity to dynamically select optimal linguistic pathways and significantly improve multilingual reasoning across diverse task domains while maintaining performance in high-resource language setting. The effectiveness of using Indonesian as a thinking language is evident in completing Pantun, where reasoning in Indonesian yields responses that better follow its traditional structure. This shows that for culturally specific tasks, low-resource languages can sometimes enable superior reasoning. We provide more specific examples, error and limitation analyses in Appendix A.7.

4.4 Interpretability Study of AdaMCOT

Prior research suggests that the reasoning processes of multilingual LLMs primarily occur within a shared, language-agnostic latent space. This shared space allows models to perform reasoning tasks across different languages, depending on the quality of their multilingual alignment. These reasoning dynamics are especially prominent in the middle and upper layers of the model, closer to the output layer (Zhao et al. 2024; Schut, Gal, and Farquhar 2025; Wang et al. 2025; Fierro et al. 2025). Furthermore, studies on knowledge neurons and cross-lingual activation patterns indicate that stronger multilingual alignment improves factual transfer and consistency (Wang et al. 2024b; Tang et al. 2024; Cao et al. 2024). To better understand how AdaMCOT enhances multilingual reasoning, we conduct an interpretability analysis using the Logit Lens (nostalgebraist 2020) and UMAP (McInnes, Healy, and Melville 2020).

Logit Lens Analysis. We use the Logit Lens to examine the model's layer-wise hidden states by projecting them onto the output vocabulary, revealing its predictive focus at each decoding step. In Figures 3a and 3b, each row represents a model layer and each column a decoding step; only the token with the highest logit is visualized to trace dominant predictions. To illustrate the impact of cross-lingual reasoning, we analyze responses to: *"Which countries have longer working hours than Mexico?"*. As shown in Figure 4, the baseline LLaMA3.1-8B-Instruction model hallucinates facts when reasoning directly in Chinese, likely due to insufficient Chinese training data.

By contrast, the AdaMCOT-enhanced model first reasons in English (Figure 3a). Although early predictions are noisy, it ultimately converges on the correct answer. It then generates the final response in Chinese, grounded in the English chain-of-thought (Figure 3b). This shift leads to more confident and coherent token selection, demonstrating that routing reasoning through high-resource languages enhances factual accuracy and reduces uncertainty.

UMAP Embedding Analysis. To further investigate cross-lingual semantic alignment, we apply UMAP to visualize the model's language-specific embedding distributions before and after applying AdaMCOT. We observe that AdaMCoT brings language-specific embeddings closer, particularly around the English centroid. For instance, in the LLaMA3.1-8B model, the average distance from non-English clusters to the English centroid decreases from 9.87 to 9.39 after applying AdaMCOT. Importantly, this alignment occurs without significant distortion of the original embedding space. As all training data are generated by the baseline model, this shift reflects AdaMCOT's ability to promote better multilingual coherence without catastrophic forgetting. Similar improvements are observed in Qwen2.5 models, confirming the generalizability of the effect. Detailed UMAP and Logit Lens plots are available in Appendix A.4.

5 Conclusion

We presented AdaMCOT, a novel framework for enhancing cross-lingual factual reasoning in LLMs via adaptive chain-of-thought prompting. By dynamically routing reasoning through strategically chosen "thinking languages," AdaMCOT mitigates performance disparities across languages, especially improving outcomes in low-resource settings, while preserving or improving accuracy in high-resource ones. Our method introduces a reward-based training procedure that selects optimal reasoning pathways using a strong LLM-based evaluator, enabling the model to learn when and how to reason cross-lingually. Comprehensive evaluations across four multilingual benchmarks demonstrate consistent gains in both factual accuracy and cross-lingual consistency. Finally, our interpretability analysis using Logit Lens and UMAP reveals that AdaMCOT promotes better semantic alignment across languages and reduces reasoning hallucinations. These findings suggest that adaptive language routing is a promising direction for improving multilingual LLM performance without requiring additional pretraining or translation-heavy pipelines.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore. The research is also supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-004). This Research is partially supported with Cloud TPUs from Google's TPU Research Cloud (TRC).

References

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

Cao, P.; Chen, Y.; Jin, Z.; Chen, Y.; Liu, K.; and Zhao, J. 2024. One Mind, Many Tongues: A Deep Dive into Language-Agnostic Knowledge Neurons in Large Language Models. *arXiv preprint arXiv:2411.17401*.

Chai, L.; Yang, J.; Sun, T.; Guo, H.; Liu, J.; Wang, B.; Liang, X.; Bai, J.; Li, T.; Peng, Q.; et al. 2025. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23550–23558.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dubois, Y.; Galambosi, B.; Liang, P.; and Hashimoto, T. B. 2024. Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators. *arXiv preprint arXiv:2404.04475*.

Fierro, C.; Foroutan, N.; Elliott, D.; and Søgaard, A. 2025. How Do Multilingual Language Models Remember Facts? arXiv:2410.14387.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. arXiv:2009.03300.

Huang, H.; Tang, T.; Zhang, D.; Zhao, W. X.; Song, T.; Xia, Y.; and Wei, F. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Huzaifah, M.; Zheng, W.; Chanpaisit, N.; and Wu, K. 2024. Evaluating Code-Switching Translation with Large Language Models. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 6381–6394. Torino, Italia: ELRA and ICCL.

Lai, V. D.; Nguyen, C. V.; Ngo, N. T.; Nguyen, T.; Dernoncourt, F.; Rossi, R. A.; and Nguyen, T. H. 2023. Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback. arXiv:2307.16039.

Li, C.; Wang, S.; Zhang, J.; and Zong, C. 2024a. Improving In-context Learning of Multilingual Generative Language Models with Cross-lingual Alignment. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8058–8076. Mexico City, Mexico: Association for Computational Linguistics.

Li, H.; Koto, F.; Wu, M.; Aji, A. F.; and Baldwin, T. 2023. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Li, J.; Huang, S.; Ching, A.; Dai, X.; and Chen, J. 2024b. PreAlign: Boosting Cross-Lingual Transfer by Early Establishment of Multilingual Alignment. *arXiv preprint arXiv:2407.16222*.

Liu, H.; Liu, J.; Cui, L.; Teng, Z.; Duan, N.; Zhou, M.; and Zhang, Y. 2023. LogiQA 2.0—An Improved Dataset for Logical Reasoning in Natural Language Understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 2947–2962.

McInnes, L.; Healy, J.; and Melville, J. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426.

nostalgebraist. 2020. Interpreting GPT: The Logit Lens. https://www.alignmentforum.org/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. Accessed: 2025-07-19.

Qi, J.; Fernández, R.; and Bisazza, A. 2023. Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10650–10666. Singapore: Association for Computational Linguistics.

Qin, L.; Chen, Q.; Wei, F.; Huang, S.; and Che, W. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *arXiv preprint arXiv:2310.14799*.

Schut, L.; Gal, Y.; and Farquhar, S. 2025. Do Multilingual LLMs Think In English? arXiv:2502.15603.

Singh, S.; et al. 2024. Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11521–11567. Bangkok, Thailand: Association for Computational Linguistics.

Tan, B. C. Z.; Weihua, Z.; Liu, Z.; Chen, N. F.; Lee, H.; Choo, K. T. W.; and Lee, R. K.-W. 2025. BLEnD-Vis: Benchmarking Multimodal Cultural Understanding in Vision Language Models. arXiv:2510.11178.

Tang, H.; Deshpande, A.; and Narasimhan, K. 2022. ALIGN-MLM: Word Embedding Alignment is Crucial for Multilingual Pre-training. *arXiv preprint arXiv:2211.08547*.

Tang, T.; Luo, W.; Huang, H.; Zhang, D.; Wang, X.; Zhao, X.; Wei, F.; and Wen, J.-R. 2024. Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5701–5715. Bangkok, Thailand: Association for Computational Linguistics.

Team, G. 2024. Gemma 2: Improving Open Language Models at a Practical Size. arXiv:2408.00118.

Team, Q. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Teknium. 2023. OpenHermes 2.5: An Open Dataset of Synthetic Data for Generalist LLM Assistants.

Wang, B.; Liu, Z.; Huang, X.; Jiao, F.; Ding, Y.; Aw, A.; and Chen, N. F. 2024a. SeaEval for Multilingual Foundation Models: From Cross-Lingual Alignment to Cultural Reasoning. arXiv:2309.04766.

Wang, M.; Adel, H.; Lange, L.; Liu, Y.; Nie, E.; Strötgen, J.; and Schütze, H. 2025. Lost in Multilinguality: Dissecting Cross-lingual Factual Inconsistency in Transformer Language Models. arXiv:2504.04264.

Wang, W.; Haddow, B.; Wu, M.; Peng, W.; and Birch, A. 2024b. Sharing matters: Analysing neurons across languages and tasks in llms. *arXiv preprint arXiv:2406.09265*.

Weihua, Z.; Lee, R. K.-W.; Liu, Z.; Kui, W.; Aw, A.; and Zou, B. 2025. CCL-XCoT: An Efficient Cross-Lingual Knowledge Transfer Method for Mitigating Hallucination Generation. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Findings of the Association for Computational Linguistics: EMNLP 2025*, 1768–1788. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-335-7.

Wendler, C.; Veselovsky, V.; Monea, G.; and West, R. 2024. Do llamas work in english? on the latent language of multilingual transformers. *arXiv preprint arXiv:2402.10588*.

Xu, H.; Kim, Y. J.; Sharaf, A.; and Awadalla, H. H. 2024. A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Yang, J. 2023. Firefly(): . https://github.com/yangjianxin1/Firefly.

Yang, W.; Li, C.; Zhang, J.; and Zong, C. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.

Zhang, S.; Fang, Q.; Zhang, Z.; Ma, Z.; Zhou, Y.; Huang, L.; Bu, M.; Gui, S.; Chen, Y.; Chen, X.; et al. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *arXiv preprint arXiv:2306.10968*.

Zhang, Y.; Chen, Q.; Li, M.; Che, W.; and Qin, L. 2024a. AutoCAP: Towards automatic cross-lingual alignment planning for zero-shot chain-of-thought. *arXiv preprint arXiv:2406.13940*.

Zhang, Y.; Wang, Y.; Liu, Z.; Wang, S.; Wang, X.; Li, P.; Sun, M.; and Liu, Y. 2024b. Enhancing Multilingual Capabilities of Large Language Models through Self-Distillation from Resource-Rich Languages. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11189–11204. Bangkok, Thailand: Association for Computational Linguistics.

Zhao, Y.; Zhang, W.; Chen, G.; Kawaguchi, K.; and Bing, L. 2024. How do Large Language Models Handle Multilingualism? *arXiv preprint arXiv:2402.18815*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2024. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Zheng, W.; Liu, Z.; Chakraborty, T.; Xu, W.; Gao, X.; Tan, B. C. Z.; Zou, B.; Liu, C.; Hu, Y.; Xie, X.; Yi, X.; Yao, J.; Wang, C.; Li, L.; Liu, R.; Liu, H.; Inoue, K.; Sumida, R.; Kawahara, T.; Xu, F.; Ye, L.; Tian, W.; Kim, D.; Jung, J.; Seo, J.; Wangsajaya, N. Y.; Duc, P. M.; Saxena, O.; Nandi, P.; Tao, X.; Karlina, W.; Luong, T.; Vasan, K. A.; Lee, R. K.-W.; and Chen, N. F. 2025. MMA-ASIA: A Multilingual and Multimodal Alignment Framework for Culturally-Grounded Evaluation. arXiv:2510.08608.

Zhong, C.; Cheng, F.; Liu, Q.; Jiang, J.; Wan, Z.; Chu, C.; Murawaki, Y.; and Kurohashi, S. 2024. Beyond English-Centric LLMs: What Language Do Multilingual Language Models Think in? *arXiv preprint arXiv:2408.10811*.

Zhu, W.; Huang, S.; Yuan, F.; She, S.; Chen, J.; and Birch, A. 2024. Question translation training for better multilingual reasoning. *arXiv preprint arXiv:2401.07817*.

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