Unlocking Multilingual Reasoning Capabilities of LLMs and LVLMs via Representation Engineering

Qiming Li[1], **Xiaocheng Feng**[1,2*], **Yixuan Ma**[1], **Ruihan Chen**[1], **Zihe Tong**[1], **Zekai Ye**[1]
Xiachong Feng[3], **Libo Qin**[4], **Haoyu Ren**[5], **Kun Chen**[5], **Yunfei Lu**[5], **Dandan Tu**[5], **Bing Qin**[1,2]
[1]Harbin Institute of Technology [2]Peng Cheng Laboratory [3]The University of Hong Kong
[4]Harbin Institute of Technology, Shenzhen [5]Huawei Technologies Co., Ltd
qmli@ir.hit.edu.cn

Abstract

Large Language Models (LLMs) and Large Vision-Language Models (LVLMs) demonstrate strong reasoning capabilities, yet their performance in English significantly outperforms that in low-resource languages, raising fairness concerns in multilingual applications. Existing approaches either rely on costly multilingual training or employ prompting with external translation tools, both of which are resource-intensive and sensitive to translation quality. To address these limitations, we propose a training-free inference-time method to enhance **M**ultilingual **R**easoning capabilities via **R**epresentation **E**ngineering (**MRRE**) without using any additional training data or tools. **MRRE** sequentially injects two precomputed vectors at specific layers during inference processing: cross-lingual reasoning enhancement vectors, which steer non-English reasoning representations toward English space to unlock multilingual reasoning, and target-language output anchoring vectors, which restore the distribution of the target language to preserve input–output language consistency. Comprehensive experiments across six advanced LLMs and LVLMs on four reasoning benchmarks demonstrate MRRE consistently enhances non-English reasoning by an average gain of 5.48% and up to 7.54% in low-resource languages (e.g., Thai and Swahili), while improving input-output language consistency by 3.78%.

1 Introduction

With the rapid development of Large Language Models (LLMs) and Large Vision-Language Models (LVLMs), foundational models such as Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-VL-7B-Instruct (Team, 2025) have demonstrated impressive capabilities in complex reasoning. However, high-resource languages such as English exhibit substantially stronger reasoning

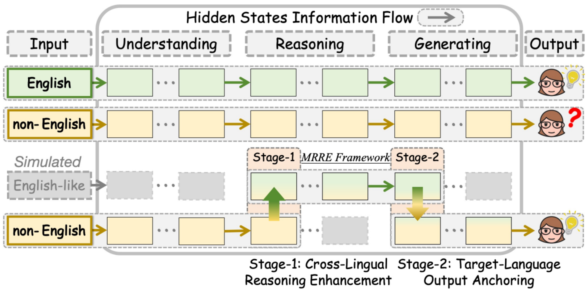


Figure 1: MRRE adopts a two-stage intervention strategy to unlock multilingual reasoning capabilities.

capabilities compared to low-resource languages, raising concerns about fairness in multilingual applications under global deployment. To address the above issues, prior works primarily focus on two directions to enhance non-English reasoning capabilities: (1) Data-driven training methods, which align multilingual embeddings (Arora et al., 2024) or construct multilingual reasoning datasets for instruction tuning (Fan et al., 2025), but inevitably depend on expensive data and incur considerable computational costs. (2) Prompting-based methods, which rely on external translation tools or models (Khandelwal et al., 2024; Liu et al., 2024), but are highly sensitive to translation quality and prompt design, accompanied by high latency. Moreover, existing methods are rarely effective across both LLMs and LVLMs. Therefore, establishing a unified and efficient paradigm for enhancing multilingual reasoning across both LLMs and LVLMs remains further exploration.

Prior studies (Zhao et al., 2024; Tang et al., 2025; Qin et al., 2025; Li et al., 2025a,b; Zhao et al., 2025) revealed the internal mechanism of multilingual reasoning: hidden states are transformed into high-resource language representations (e.g., English) in early layers, then exploited for reasoning from middle to later layers, and finally restored target language features in late layers. However, as shown in Figure 2, we observe that hidden

*Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 24810–24829
July 2-7, 2026

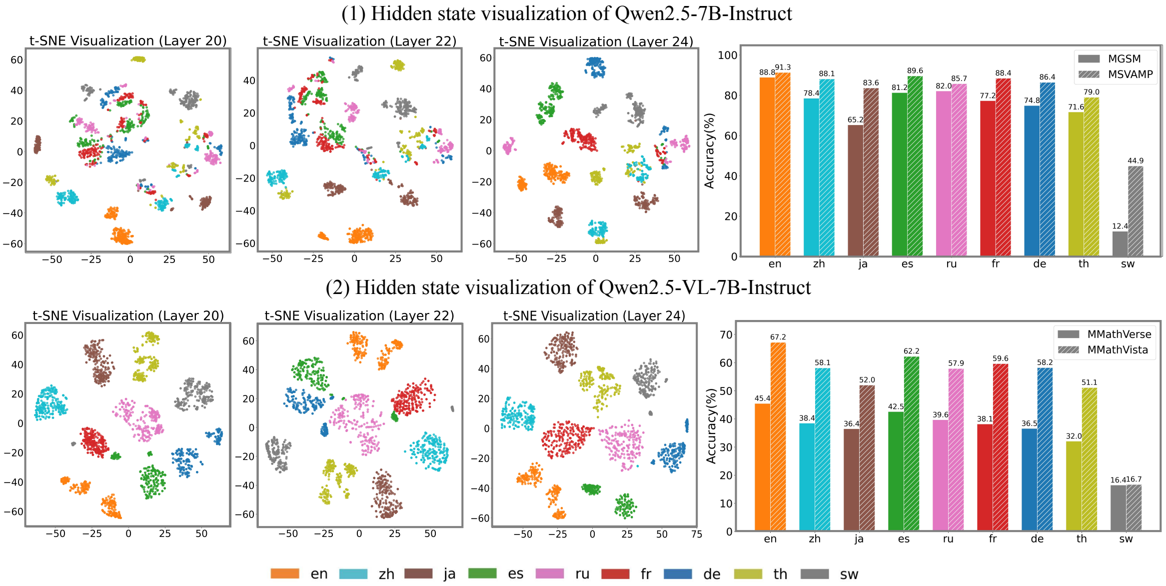


Figure 2: t-SNE hidden state visualization and reasoning performance of Qwen2.5-7B-Instruct and Qwen2.5-VL-7B-Instruct. The reasoning capability in English is substantially stronger than in other languages.

states in reasoning-related layers of both LLMs and LVLMs still exhibit significant differences between English and non-English inputs. This consistent cross-architecture phenomenon motivates us to steer the distribution of non-English hidden states toward English, endowing both LLMs and LVLMs with English-level reasoning capabilities under a general framework. As shown in Figure 1, we propose a training-free inference-time method to enhance **M**ultilingual **R**easoning capabilities via **R**epresentation **E**ngineering (**MRRE**) without using any additional training data or tools. **MRRE** sequentially injects two precomputed vectors at specific layers during forward passing: ***cross-lingual reasoning enhancement vectors***, which steer non-English reasoning representations toward English distributions to strengthen reasoning, and ***target-language output anchoring vectors***, which restore the distribution of the target language to preserve input–output language consistency.

Experimental results of six advanced LLMs and LVLMs on four reasoning benchmarks demonstrate that MRRE enhances non-English reasoning by an average gain of 5.48% and up to 7.54% in low-resource languages (Thai & Swahili), while improving input-output language consistency by 3.78%.

2 Related Works

2.1 Multilingual Foundation Models

To address the multilingual demands of real-world global applications, recent studies have extended foundation models to multilingual models. Advanced LLMs like Qwen3-8B (Yang et al., 2025) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024), are pretrained and instruction-tuned on massive multilingual data, enabling them to respond across diverse languages. Furthermore, advanced LVLMs like Qwen2.5-VL-7B-Instruct (Team, 2025) and InternVL3.5-8B-Chat (Wang et al., 2025a), integrate visual encoders with these language backbones, thereby inheriting comparable multilingual capabilities. However, the unbalanced distribution of training data across languages results in a significant performance disparity between high-resource languages (e.g., English) and low-resource language (e.g., Swahili), raising concerns of fairness.

2.2 Multilingual Reasoning Enhancement

To enhance multilingual reasoning, previous research can be divided into two categories: **(1) Data-driven training** methods enhance multilingual reasoning by aligning cross-lingual representations or fine-tuning with multilingual supervision, including contrastive alignment (Li et al., 2023; Huang et al., 2024; Arora et al., 2024) and reasoning-specific instruction tuning (Zhang et al., 2024a; Geng et al., 2024; Lai and Nissim, 2024; Fan et al., 2025). Although effective, these approaches incur substantial data and computational costs. **(2) Prompting** methods mitigate language imbalance without parameter updates, including through direct multilingual inputs (Sakai et al., 2024; Khandelwal et al., 2024), pivot-language translation

(Liu et al., 2024), or chain-of-thought prompting (Wang et al., 2024; Qin et al., 2023), although effectiveness remain sensitive to translation quality and prompt design, accompanying by high latency. In contrast to previous work, MRRE is the first training-free inference-time method via representation engineering without using any additional data or tools, which clearly distinguishes MRRE from existing methods.

3 Methods

3.1 Task Formulation

We restrict our scope to models that are based on auto-regressive Transformer architecture (Vaswani et al., 2017), as it is adopted by most SOTA LLMs and LVLMs. The input sequence of LLMs and LVLMs are processed through L transformer layers of the language decoder, each consisting of multi-head self-attention (MHSA), feed-forward network (FFN) that is usually a multilayer perception (MLP), and a residual stream is applied between each component. The hidden state $\boldsymbol{h}^{(l)} \in \mathbb{R}^d$ for token t at layer l under input sequence $R(x)$ can be computed from the previous layer:

$$\boldsymbol{h}^{(l)}(R(x), t) = \boldsymbol{h}^{(l-1)}(R(x), t) + \boldsymbol{a}^{(l)} + \boldsymbol{m}^{(l)}, \tag{1}$$

where $\boldsymbol{a}^{(l)}$ and $\boldsymbol{m}^{(l)}$ are the outputs of the MHSA and FFN component at layer l. Finally, the model predicts the next token in an auto-regressive manner based on the hidden state of the last layer.

In this paper, to bridge the significant performance gap between English and non-English languages in reasoning tasks, we propose a representation engineering approach that applies a two-stage hierarchical intervention on the hidden states of specific layers within the language decoder, unlocking non-English reasoning capabilities while preserving input-output language consistency.

3.2 Cross-Lingual Reasoning Enhancement

Prior studies (Zhao et al., 2024; Tang et al., 2025) have shown that the mid-layer hidden states play a critical role in shaping reasoning latent representation. Furthermore, as shown in Figure 2, we observe that the hidden states from middle to deeper layers under different languages exhibit significant differences in the latent space across both LLMs and LVLMs. These findings motivate us to propose a representation engineering strategy that aligns the reasoning capability of non-English with English, which exhibits stronger reasoning performance.

Since LLMs and LVLMs generate tokens in an auto-regressive manner, we focus on the hidden state of the last token, which aggregates the most comprehensive visual and textual information. To precisely estimate the reasoning enhancement direction for hidden states, we define ***cross-lingual reasoning enhancement vectors***, which align the hidden states of the reasoning chain in the target non-English language with the stronger reasoning chain in English. These vectors are computed by comparing the output last token's hidden states on a set of reasoning problems $\mathcal{X}$. For each problem $x \in \mathcal{X}$, we feed the model with both English prompt and parallel non-English prompt, generating the corresponding reasoning responses, $R_{\mathcal{E}}(x)$ and $R_{\mathcal{T}}(x)$. We define a difference vector $\Delta\boldsymbol{h}_x^{(l)}$ to estimate the latent difference between English and non-English:

$$\Delta\boldsymbol{h}_x^{(l)} = \boldsymbol{h}^{(l)}(R_{\mathcal{E}}(x), t_{\text{last}}) - \boldsymbol{h}^{(l)}(R_{\mathcal{T}}(x), t_{\text{last}}), \tag{2}$$

The ***cross-lingual reasoning enhancement vector*** at l-th layer, $\boldsymbol{v}_r^{(l)}$, is then computed as the mean of these difference vectors over the entire set $\mathcal{X}$:

$$\boldsymbol{v}_r^{(l)} = \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} \Delta\boldsymbol{h}_x^{(l)}. \tag{3}$$

By steering the target-language hidden states along $\boldsymbol{v}_r^{(l)}$, we shift the non-English reasoning chain to a stronger reasoning chain shaped by English. This intervention enables the model to exhibit English-level reasoning capability when answering questions in the non-English language.

3.3 Target-Language Output Anchoring

However, merely applying the ***cross-lingual enhancement vectors*** may compromise input-output language consistency. When the steered hidden states pass through the layers responsible for constructing the output language representation, the models process them as if they were English inputs, resulting in English outputs rather than the target language. To address this issue, we propose ***target-language output anchoring vectors***, which guide the model to steer the English-like output distribution toward the target non-English language in the layers responsible for constructing the output language representation, thereby ensuring input-output language consistency. The ***target-language output anchoring vector*** at l'-th layer, $\boldsymbol{v}_a^{(l')}$, is computed by taking the difference between the last token representations of two fixed language forcing

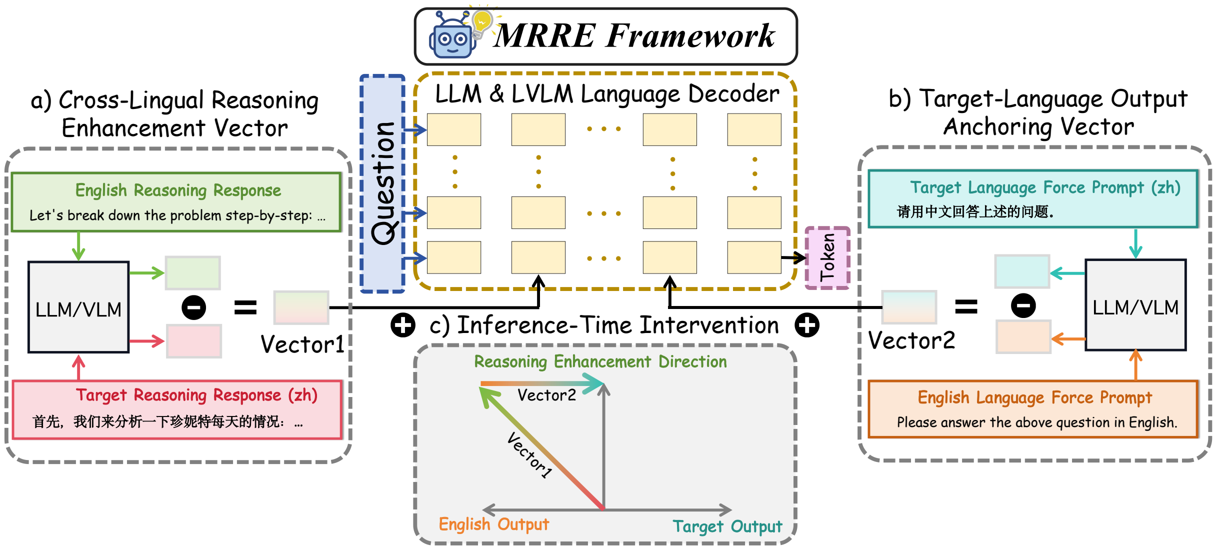


Figure 3: An overview of our proposed MRRE method. Each rectangle represents the model's hidden state during the forward passing. MRRE consists of three key stages: **a) Cross-Lingual Reasoning Enhancement Vectors** §3.2 are derived from the hidden state differences between English and non-English reasoning responses. **b) Target-Language Output Anchoring Vectors** §3.3 are derived from the hidden state differences between non-English and English language forcing prompts. **c) Hierarchical Inference-Time Intervention** §3.4: Precomputed vectors are sequentially injected into the last-token representations at specific layers during forward passing, thereby enhancing non-English reasoning capabilities while preserving input-output language consistency.

prompts: $P_{\mathcal{T}}$ ("Please answer this question in <*target language*>" translated into <*target language*>), and $P_{\mathcal{E}}$ ("Please answer this question in English"):

$$\boldsymbol{v}_a^{(l')} = \boldsymbol{h}^{(l')}(P_{\mathcal{T}}, t_{\text{last}}) - \boldsymbol{h}^{(l')}(P_{\mathcal{E}}, t_{\text{last}}). \quad (4)$$

This vector provides a precise estimate of the hidden state distribution shift from English to the target language, enabling the model to recover output target-language distribution in later layers and ensuring input-output language consistency.

3.4 Hierarchical Inference-Time Intervention

Considering the autoregressive decoding mechanism of language models, we propose a two-stage, hierarchical inference-time intervention method. First we apply ***cross-lingual enhancement vectors*** to the last token of the current reasoning response at middle layer, steering non-English hidden state $\boldsymbol{h}_t^{(l)}$ to an English-like hidden state $\tilde{\boldsymbol{h}}^{(l)}$:

$$\hat{\boldsymbol{h}}^{(l)} = \boldsymbol{h}^{(l)} + \alpha_1 \cdot \tilde{\boldsymbol{v}}_r^{(l)}, \tilde{\boldsymbol{h}}^{(l)} = \hat{\boldsymbol{h}}^{(l)} \cdot \frac{\|\boldsymbol{h}^{(l)}\|_2}{\|\hat{\boldsymbol{h}}^{(l)}\|_2}, \quad (5)$$

where α_1 denotes scaling coefficients and $||\cdot||$ represents the ℓ_2 norms of the activation vectors. The normalization strategy ensures that the vector scale remains consistent before and after intervention, preventing undesired magnitude shifts that may distort downstream representations.

The model exhibits English-level reasoning capability from middle to later layers as if processing an English problem, producing English output hidden states $\boldsymbol{h}^{(l')}$ at l'-th layer. Then we apply the ***target-language output anchoring vectors*** to steer English output hidden state back to the target $\tilde{\boldsymbol{h}}^{(l')}$:

$$\hat{\boldsymbol{h}}^{(l')} = \boldsymbol{h}^{(l')} + \alpha_2 \cdot \tilde{\boldsymbol{v}}_a^{(l')}, \tilde{\boldsymbol{h}}^{(l')} = \hat{\boldsymbol{h}}^{(l')} \cdot \frac{\|\boldsymbol{h}^{(l')}\|_2}{\|\hat{\boldsymbol{h}}^{(l')}\|_2}, \quad (6)$$

Finally, the newly generated token is then appended to the current input sequence. In the middle layers, the hidden state of updated input sequence can be *"translated"* into English, activating stronger reasoning capabilities under English-like states; and in later layers, it can be *"back-translated"* to the target language, ensuring input-output language consistency. This information flow continues until end-of-sequence token is produced.

4 Experiments

4.1 Experimental Setup

To thoroughly evaluate our proposed MRRE's performance and generalization across models and languages, we construct systematic experiments.

Baseline Models. We evaluate our proposed MRRE method on six SOTA models to demonstrate its broad applicability across LLM and LVLMs.

Model	En	Zh	Ja	Es	Ru	Fr	De	Th	Sw	LC	En	Zh	Ja	Es	Ru	Fr	De	Th	Sw	LC
LLMs					***MGSM***										***MSVAMP***					
Qwen2.5-7B-Instruct	88.8	78.4	65.2	**81.2**	82.0	77.2	74.8	71.6	12.4	84.7	91.3	88.1	83.6	89.6	85.7	88.4	86.4	79.0	44.9	86.3
+ Few-Shot	–	79.1	66.8	80.2	82.5	77.8	75.6	72.3	14.9	87.6	–	88.6	84.1	**90.1**	86.2	89.0	87.1	80.4	46.8	88.9
+ SFT	–	76.4	64.9	78.6	80.8	75.9	74.2	70.1	13.7	86.0	–	86.9	82.3	88.7	84.6	87.5	85.9	78.8	45.1	87.4
+ Multilingual-CoT	–	80.6	68.9	80.4	83.9	78.6	77.3	73.9	15.3	82.1	–	88.4	85.2	91.3	86.5	90.0	**88.2**	81.6	49.3	84.8
+ Language forcing	–	77.2	67.6	79.2	82.0	76.0	76.8	71.6	16.0	92.2	–	86.2	82.0	88.8	85.1	87.8	86.4	79.6	45.7	89.4
+ MRRE	–	**81.6**	**71.2**	**81.2**	**84.7**	**79.2**	**78.4**	**74.8**	**17.2**	**92.7**	–	**88.7**	**84.7**	90.0	**86.5**	**90.1**	87.4	**82.6**	**52.3**	**90.3**
Qwen3-8B	90.8	80.4	76.4	83.2	88.4	82.4	84.0	85.6	40.4	89.3	91.9	88.1	87.4	90.7	87.0	90.1	89.1	83.1	67.8	89.4
+ Language forcing	–	82.0	76.8	84.0	**88.8**	79.6	82.0	84.0	37.6	93.2	–	88.6	86.8	90.0	87.7	89.5	88.7	83.1	62.3	91.8
+ MRRE	–	**84.7**	**81.4**	**84.2**	**88.8**	**84.1**	**85.5**	**86.0**	**45.8**	**95.6**	–	**89.1**	**88.2**	**90.9**	**88.3**	**90.3**	**89.4**	**85.4**	**73.4**	**92.5**
Llama-3.1-8B- Instruct	77.2	61.2	38.0	64.8	53.2	63.6	67.2	54.4	44.8	95.3	78.4	65.7	53.8	71.4	56.6	76.4	72.9	55.8	50.8	94.9
+ Language forcing	–	57.6	44.8	63.6	62.4	68.0	66.4	53.2	45.6	95.1	–	58.5	59.0	71.5	**66.0**	73.0	69.9	56.5	50.5	95.1
+ MRRE	–	**66.6**	**47.7**	**68.7**	**64.5**	**70.0**	**71.2**	**64.3**	**53.9**	**96.6**	–	**70.6**	**64.3**	**73.2**	65.4	**76.4**	**74.2**	**63.9**	**57.6**	**96.4**
LVLMs					***MMathVerse***										***MMathVista***					
Qwen2.5-VL-7B-Instruct	45.4	38.4	36.4	42.5	39.6	38.1	36.5	32.0	16.4	92.5	67.2	58.1	52.0	62.2	57.9	59.6	58.2	51.1	16.7	94.1
+ Few-Shot	–	45.3	40.0	45.2	44.1	42.7	43.1	35.5	23.0	92.6	–	56.9	57.4	60.9	59.7	61.6	56.9	52.7	34.8	94.0
+ SFT	–	34.0	33.4	33.9	34.5	30.4	31.7	28.3	17.0	92.7	–	51.5	49.1	54.3	54.7	53.9	51.7	44.9	18.2	94.5
+ Multilingual-CoT	–	44.9	42.0	46.7	44.5	44.9	44.0	37.8	22.4	90.4	–	61.7	57.4	63.9	62.5	59.7	61.2	52.8	30.3	92.6
+ Language forcing	–	36.8	37.7	43.6	40.6	39.6	40.0	43.5	17.8	91.8	–	60.0	53.0	61.2	58.3	59.4	56.9	53.7	10.1	94.4
+ MRRE	–	**45.9**	**44.6**	**47.5**	**44.6**	**45.8**	**45.7**	**48.4**	**33.8**	**93.1**	–	**62.3**	**61.1**	**64.2**	**62.8**	**61.2**	**64.1**	**57.3**	**28.6**	**95.5**
LLaVA-Onevision-7B	29.9	28.9	25.8	29.2	26.2	26.8	25.3	23.5	11.3	76.9	61.9	51.1	43.7	**58.0**	49.2	55.8	48.0	41.7	16.3	75.0
+ Language forcing	–	26.6	22.7	25.9	22.9	20.3	22.0	26.3	4.0	**77.8**	–	44.6	38.6	45.1	41.1	41.8	37.1	26.7	10.8	78.4
+ MRRE	–	**32.9**	**30.0**	**30.0**	**28.2**	**30.0**	**29.5**	**26.5**	**19.1**	**77.8**	–	**52.3**	**48.2**	57.0	**55.7**	**56.6**	**54.7**	**48.3**	**32.1**	**80.3**
InternVL3.5-VL-8B-Chat	57.6	52.3	48.9	52.3	49.0	48.5	46.6	34.4	32.3	76.3	72.2	59.7	60.2	67.9	68.7	67.9	68.7	62.5	40.6	79.6
+ Language forcing	–	54.7	50.8	51.4	49.4	49.1	49.8	35.5	36.3	**84.2**	–	62.4	61.1	64.1	62.9	61.5	64.1	57.8	34.8	85.6
+ MRRE	-	**55.6**	**51.2**	**53.3**	**50.1**	**50.3**	**52.3**	**44.6**	**42.3**	81.9	–	**63.5**	**66.9**	**70.7**	**70.5**	**70.1**	**69.2**	**66.3**	**46.6**	**86.4**

Table 1: Accuracy (%) and Language Consistency (LC, %) of three advanced LLMs and three advanced LVLMs with different settings across 8 languages (Zh, Ja, Es, Ru, Fr, De, Th, Sw) and 4 reasoning benchmarks: *MGSM, MSVAMP, MMathVerse, MMathVista*. Best performances for each experimental settings are **bolded**.

Model	*MMathVerse*						*MMathVista*		
	T-D	**T-L**	**V-I**	**V-D**	**V-O**	**Overall**	**General**	**Math**	**Overall**
Qwen2.5-VL-7B	39.9	35.1	32.1	32.5	34.6	34.9	53.5	50.7	52.0
+ Few-Shot	46.2	40.4	37.8	37.2	37.8	39.9	55.0	55.2	55.1
+ SFT	35.6	31.6	28.5	28.2	28.1	30.4	42.8	48.0	45.8
+ Multilingual-CoT	47.3	41.6	39.9	39.9	40.3	41.8	57.0	55.5	56.2
+ Language forcing	44.0	37.2	34.9	35.1	37.0	37.7	52.6	50.8	51.6
+ MRRE	**51.2**(+11.3)	**45.8**(+10.7)	**42.2**(+10.1)	**41.2**(+8.7)	**41.9**(+7.3)	**44.5**(+9.6)	**55.5**(+2.0)	**58.5**(+7.8)	**57.1**(+5.1)
LLaVA-Onevision-7B	27.5	28.4	24.8	22.6	**22.4**	24.6	44.7	46.1	45.5
+ Language forcing	28.4	25.7	24.2	19.8	8.5	29.3	21.3	41.0	35.8
+ MRRE	**37.4**(+9.9)	**32.1**(+3.7)	**30.9**(+6.1)	**28.3**(+5.7)	21.2(-1.2)	**30.0**(+5.4)	**48.6**(+3.9)	**52.3**(+6.2)	**50.6**(+5.1)
InternVL3.5-8B	48.3	41.4	33.3	39.4	37.5	40.1	62.0	61.0	61.5
+ Language forcing	47.8	40.3	36.1	38.1	41.7	40.8	60.8	61.8	61.4
+ MRRE	**53.0**(+4.7)	**47.2**(+5.8)	**41.7**(+8.4)	**44.3**(+4.9)	**48.0**(+10.5)	**46.8**(+6.7)	**65.3**(+3.3)	**64.0**(+3.0)	**63.8**(+2.3)

Table 2: Mean performance (%) of three advanced LVLMs with diverse experimental settings across 8 languages (Zh, Ja, Es, Ru, Fr, De, Th, Sw) and 2 benchmarks: MMathVerse (*Text-Dominant, Text-Lite, Vision-Integrated, Vision-Dominant, and Vision-Only categories, Overall*), MathVista (*General, and Math-related categories, Overall*).

LLMs: Qwen2.5-7B-Instruct (Team, 2024), Qwen3-8B (Yang et al., 2025), and Llama-3.1-8B-Instruct (Grattafiori et al., 2024).

LVLMs: Qwen2.5-VL-7B-Instruct (Team, 2025), LLaVA-OneVision (Li et al., 2024) and InternVL3.5-8B-Chat (Wang et al., 2025a).

Baseline Methods. We select several methods that are effective for multilingual reasoning tasks in both LLMs and VLMs, including Few-Shot, SFT, Multilingual-CoT (Barua et al., 2025), and Language forcing (Wang et al., 2025b).

Benchmarks. We select four challenging benchmarks to evaluate multilingual reasoning of both LLMs and LVLMs, covering both mathematical and general reasoning capabilities, including **MGSM** (Shi et al., 2022), **MSVAMP** (Chen et al., 2023), **MMathVerse** (Zhang et al., 2024b) and **MMathVista** (Lu et al., 2023).

We adopt *Accuracy* as the evaluation metric. Language Consistency (LC) denotes the proportion of responses generated in the target language.

Implementation Details. We randomly sample 100 instances from MGSM and MathVerse to construct MRRE vectors for LLMs and LVLMs, respectively. All experiments are conducted on 8 $\times$ NVIDIA A100 80GB. See Appendix B for details.

4.2 Main Results

Based on experimental results presented in Table 1 & 2, we can draw the following key conclusions:

(1) Effective multilingual reasoning enhancement performance. MRRE achieves effective performance across high-resource and low-resource non-English languages on LLMs and LVLMs, leading to an average improvement of 5.48% on four benchmarks, **while improving input-output language consistency by 3.78%.** Notably, improvements are significantly enhanced in low-resource languages (Th, Sw) by 7.54%.

(2) Multimodal generalizability. MRRE consistently improves performance across purely textual and vision–language reasoning tasks. Notably, results on MMathVerse demonstrate that MRRE enhances reasoning in all five categories, including *Text-Dominant*, *Text-Lite*, *Vision-Integrated*, *Vision-Dominant*, and *Vision-Only* tasks, proving MRRE's effectiveness regardless of modality composition.

(3) Cross-model generalizability. By systematically evaluating diverse LLM and LVLM backbones, we find that MRRE does not rely on a particular model architecture. In contrast, MRRE benefits from universal cross-lingual difference in hidden states, ensuring broad application across other SOTA open-source models.

(4) Cross-dataset generalizability. Although the construction of MRRE vectors relies on reasoning data from MGSM and MMathVerse, these vectors remain effective on out-of-distribution datasets such as MSVAMP and MMathVista, suggesting that the intervention represents a generalizable direction rather than being merely tailored to a specific dataset. Moreover, the improvement on MMathVista *General* subset indicates that MRRE not only enhances mathematical reasoning, but also strengthens general reasoning capabilities.

5 Analysis and Discussions

5.1 Analysis of Intervention

To thoroughly visualize the effect of MRRE on hidden states, we conduct a stepwise analysis along

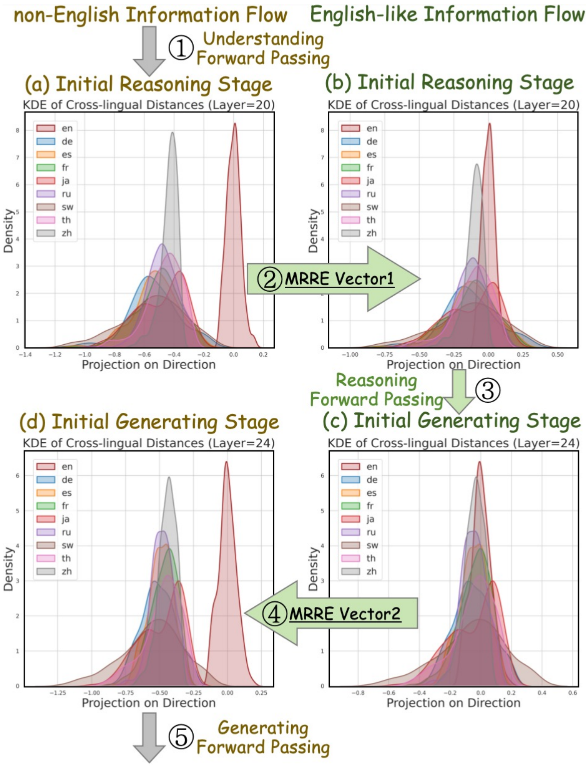


Figure 4: Kernel Density Estimate (KDE) visualization plots of cross-lingual hidden states within Qwen2.5-7B-Instruct before and after two types of intervention. The x-axis represents the SVM-derived signed distance to the mean English representation; and the y-axis represents the estimated probability density.

the information flow of token generation: ① the model first receives non-English input and undergoes understanding forward passing until the initial reasoning stage. As shown in Figure 4 (a), we observe a significant difference between English and non-English representations in the latent space. Moreover, since Qwen2.5-7B-Instruct is trained on large amounts of English and Chinese data, the hidden states of these two high-resource languages exhibit much more densely clustered than others. ② Under the intervention of the ***cross-lingual enhancement vectors***, as shown in Figure 4 (b) non-English hidden states shift closer to English. ③ The English-like states then undergo reasoning forward passing until initial generating stage critical for language outputs. At this stage, the model reasons as if it were processing an English query, ultimately producing English-like representations. As illustrated in Figure 4 (c), English and non-English hidden states become highly aligned. ④ Under the intervention of the ***target-language output anchoring vectors***, as shown in Figure 4 (d), the hidden states restore the standard output distribution of the

Model	Zh	Ja	Es	Ru	Fr	De	Th	Sw
Qwen2.5-7B as *Language Backbone*								
Qwen2.5-VL-7B	58.1	52.0	62.2	57.9	59.6	58.2	51.1	16.7
+ MRRE *Vanilla*	**62.3**	**61.1**	**64.2**	**62.8**	61.2	**64.1**	**57.3**	28.6
+ MRRE *L-B*	58.9	55.7	62.0	59.6	**61.4**	62.7	54.9	**33.1**
Qwen3-8B as *Language Backbone*								
InternVL3.5-8B	59.7	60.2	67.9	68.7	67.9	68.7	62.5	40.6
+ MRRE *Vanilla*	**63.5**	**66.9**	**70.7**	**70.5**	**70.1**	69.2	**66.3**	46.6
+ MRRE *L-B*	61.2	62.1	68.5	69.0	68.8	**69.4**	64.5	**49.4**

Table 3: Performance (%) of the MathVista benchmark across languages. *Vanilla* represents standard MRRE method, and *L-B* represents MRRE method using vectors from responding LVLM's *Language Backbone*.

Model	Zh	Ja	Es	Ru	Fr	De	Th	Sw
LLM on MGSM								
Qwen2.5-7B	78.4	65.2	**81.2**	82.0	77.2	74.8	71.6	12.4
+ MRRE *Vanilla*	81.6	**71.2**	**81.2**	**84.7**	**79.2**	**78.4**	74.8	17.2
+ MRRE *Debias*	**83.2**	68.8	**81.2**	83.6	77.2	77.2	**76.4**	**19.2**
LVLM on MMathVista								
Qwen2.5-VL-7B	58.1	52.0	62.2	57.9	59.6	58.2	51.1	16.7
+ MRRE *Vanilla*	**62.3**	**61.1**	**64.2**	**62.8**	**61.2**	**64.1**	**57.3**	**28.6**
+ MRRE *Debias*	54.3	48.0	56.6	54.4	54.8	56.1	48.5	11.9

Table 4: Performance (%) of the MGSM and MMathVista benchmark across languages. *Vanilla* represents standard MRRE method, and *Debias* represents MRRE method using ***latent english debiasing vectors***.

target language, ensuring input and output language consistency. ⑤ Finally, the anchored hidden states continue forward passing until generate the next token in target language. This lifetime visualization of token generation demonstrates that MRRE achieves the foundational motivation we set forth.

5.2 Cross-modal Generalization of Vectors

In this subsection, we thoroughly explore the cross-modal generalization of MRRE vectors. LVLMs are trained by a **L**anguage **B**ackbone (*L-B*) LLM jointly with a specific visual encoder, which results in high similarity between their output hidden states in the latent space. Furthermore, current language backbone LLMs exhibit stronger reasoning capabilities than corresponding LVLMs (Chen et al., 2025). Building on these two observations, we hypothesize that cross-lingual reasoning enhancement vectors derived from language backbones can also enhance reasoning capability of LVLMs. To validate this hypothesis, we conduct experiments with Qwen2.5-VL-7B-Instruct and InternVL3.5-8B, replacing the cross-lingual reasoning enhancement vectors with those vectors derived from their language backbone, Qwen2.5-7B-Instruct and Qwen3-8B. As shown in Table 3, MRRE *L-B* consistently improves the reasoning performance across all non-English languages. Remarkably, MRRE *L-B* even surpasses MRRE *Vanilla* in French and Swahili. These findings indicate that vectors from language backbones generalize well to LVLMs, highlighting MRRE's cross-modal generalization capability. They also suggest that LVLMs inherit multilingual reasoning capabilities from their language backbones LLMs to some extent, leading to similar multilingual shift directions.

5.3 Mitigating English Bias in Latent Space

In this subsection, we further explore how to enhance multilingual reasoning capabilities during the understanding forward passing. Prior study (Zhao et al., 2024; Xiao et al., 2026) indicates that hidden states encode low-level semantic representations and exhibit strong linguistic characteristics at this stage. Since a substantial portion of the training data is in English, when the model processes non-English queries, the hidden-state distribution during forward passing tends to shift toward the default English distribution. We hypothesize these shifts in early layers may distort the encoded low-level semantic representations, leading to performance drops. To address this issue, we introduce an alternative multilingual reasoning enhancement method, which only leverages the ***cross-lingual enhancement vectors*** in a reversed direction. We denote these vectors as the ***latent english debiasing vectors***, designed to mitigate English bias in latent space during understanding forward passing:

$$\boldsymbol{v}^{(l)}_{debias} = -\frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} \Delta \boldsymbol{h}^{(l)}_{x}. \qquad (7)$$

To validate this hypothesis, we conduct experiments with Qwen2.5-7B-Instruct and Qwen2.5-VL-7B-Instruct and apply ***latent english debiasing vectors*** for each model. As illustrated in Table 4, the MRRE *Debias* method consistently improves cross-lingual performance on the Qwen2.5-7B-Instruct across all evaluated languages, and even surpasses the vanilla MRRE *Vanilla* on Chinese, Thai, and Swahili. However, when applied to the Qwen2.5-VL-7B-Instruct, MRRE *Debias* exhibits a consistent performance decline. According to prior study (Ye et al., 2025), we posit that LVLMs rely predom-

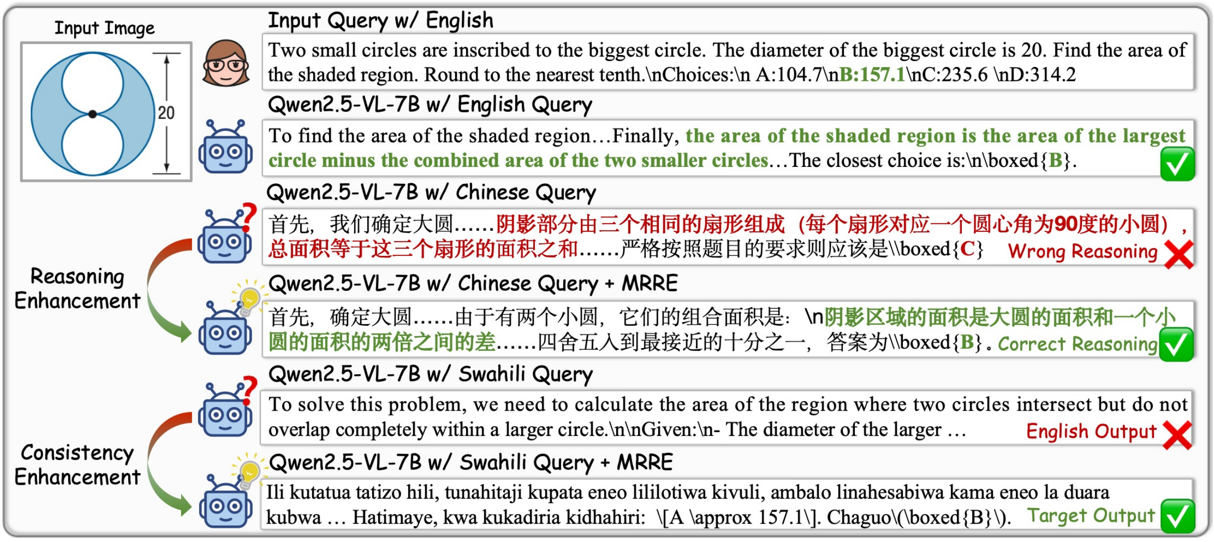


Figure 5: Case study of Qwen2.5-VL-7B-Instruct on the MathVerse benchmark.

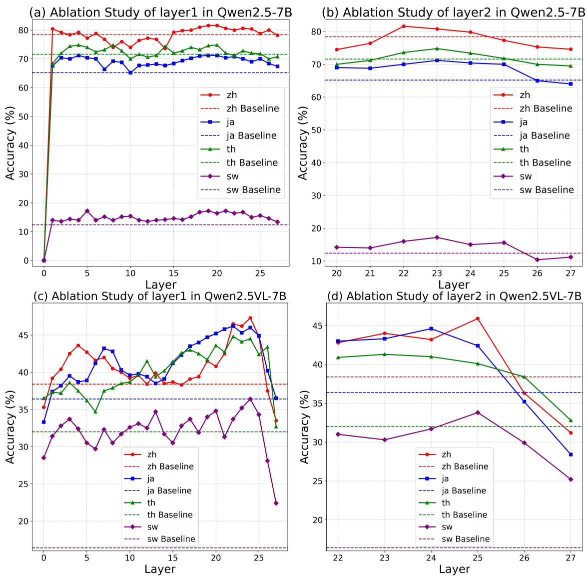


Figure 6: Analysis of layer selections of MRRE.

Model	Zh	Ja	Es	Ru	Fr	Th	Sw	LC
Qwen2.5-7B	78.4	65.2	81.2	82.0	77.2	71.6	12.4	84.7
+ MRRE *Vector1*	**83.2**	**73.7**	**82.3**	**85.2**	**80.1**	**75.2**	**19.2**	26.2
+ MRRE *Vanilla*	81.6	71.2	81.2	84.7	79.2	74.8	17.2	**92.7**
Qwen2.5VL-7B	58.1	52.0	62.2	57.9	59.6	51.1	16.7	94.1
+ MRRE *Vector1*	**63.4**	**62.3**	**64.9**	**64.2**	**63.9**	**62.5**	**35.4**	30.0
+ MRRE *Vanilla*	62.3	61.1	64.2	62.8	61.2	57.3	28.6	**95.5**

Table 5: Ablation studies of MRRE vectors.

inantly on English data during the image–text alignment stage of pre-training, which endows English with stronger multimodal representations. Consequently, injecting ***latent english debiasing vectors*** into non-English hidden states disrupts multimodal representations, leading to performance drops.

5.4 Analysis of Hyperparameters

This subsection systematically examines the influence of the intervention layer1 l and layer2 l' for two designed vectors, respectively. As shown in Figure 6, applying the first MRRE vector near layer 20 yields the best improvement in the model's reasoning ability. Given the first intervention at layer 20, applying the second vector near layer 23 achieves the next optimal boost. LVLMs exhibit a similar pattern, with the optimal layers being 22 and 24, respectively. As shown in Table 5, the second vector exhibits a trade-off strategy, which decreases little reasoning capabilities but significantly increases the consistency. Furthermore, we find MRRE achieves optimal results when $\alpha_1 = 1$ and $\alpha_2 = 0.75$. See Appendix B.3 for more ablation studies of intervention strengths α_1 and α_2.

5.5 Case Study

As shown in Figure 5, MRRE not only enhances the model's reasoning capability in non-English settings, but also improves the consistency between input and output languages. See Appendix C for fine-grained case studies across LLMs and LVLMs.

5.6 Efficiency Analysis

To assess the practical deployability of MRRE, we analyze its computational overhead.

Offline Preparation The steering vectors are precomputed using a subset of 100 samples. This involves only element-wise subtraction and normalization, representing a negligible one-time cost.

Inference Overhead During inference, MRRE modifies the hidden states via two vector additions per token. We measure the latency using Qwen2.5-7B on a single NVIDIA A100 GPU. As reported in Table 6, the impact on *Time to First Token* (TTFT) and *Time Per Output Token* (TPOT) is minimal (< 0.5ms), confirming that MRRE maintains the high-speed generation characteristic.

Table 6: Inference latency comparison of the base model and MRRE.

Method	TTFT (ms)	TPOT (ms)
Qwen2.5-7B	83.7	32.1
+ MRRE	84.1 (+0.4)	32.5 (+0.4)

6 Conclusion

We propose a training-free inference-time method to enhance **M**ultilingual **R**easoning capabilities via **R**epresentation **E**ngineering (**MRRE**), which applies cross-lingual reasoning enhancement vectors and target-language output anchoring vectors sequentially at specific layers of LLMs and LVLMs during forward passing. Comprehensive results demonstrate its effectiveness and generalizability.

Limitations

One limitation of our work is that MRRE requires access to the internal representations of the model, making it infeasible for closed-source LLMs or LVLMs. Furthermore, due to the constraints of cost and resources, we only conduct experiments on these widely used benchmarks and models.

Acknowledgements

Xiaocheng Feng is the corresponding author of this work. We thank the anonymous reviewers for their insightful comments. This work was supported by the National Natural Science Foundation of China (NSFC) (grant 62522603, 62276078), the Key R&D Program of Heilongjiang via grant 2022ZX01A32, the Fundamental Research Funds for the Central Universities (XNJKKGYDJ2024013) .

References

Gaurav Arora, Srujana Merugu, Shreya Jain, and Vaibhav Saxena. 2024. Towards robust knowledge representations in multilingual llms for equivalence and inheritance based consistent reasoning. *arXiv preprint arXiv:2410.14235*.

Josh Barua, Seun Eisape, Kayo Yin, and Alane Suhr. 2025. Long chain-of-thought reasoning across languages.

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

Shiqi Chen, Jinghan Zhang, Tongyao Zhu, Wei Liu, Siyang Gao, Miao Xiong, Manling Li, and Junxian He. 2025. Bring reason to vision: Understanding perception and reasoning through model merging. *arXiv preprint arXiv:2505.05464*.

Yuchun Fan, Yongyu Mu, Yilin Wang, Lei Huang, Junhao Ruan, Bei Li, Tong Xiao, Shujian Huang, Xiaocheng Feng, and Jingbo Zhu. 2025. Slam: Towards efficient multilingual reasoning via selective language alignment. *arXiv preprint arXiv:2501.03681*.

Xiang Geng, Ming Zhu, Jiahuan Li, Zhejian Lai, Wei Zou, Shuaijie She, Jiaxin Guo, Xiaofeng Zhao, Yinglu Li, Yuang Li, et al. 2024. Why not transform chat large language models to non-english? *arXiv preprint arXiv:2405.13923*.

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

Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. 2024. Mindmerger: Efficiently boosting llm reasoning in non-english languages. *Advances in Neural Information Processing Systems*, 37:34161–34187.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aditi Khandelwal, Utkarsh Agarwal, Kumar Tanmay, and Monojit Choudhury. 2024. Do moral judgment and reasoning capability of llms change with language? a study using the multilingual defining issues test. *arXiv preprint arXiv:2402.02135*.

Huiyuan Lai and Malvina Nissim. 2024. mcot: Multilingual instruction tuning for reasoning consistency in language models. *arXiv preprint arXiv:2406.02301*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2023. Improving in-context learning of multilingual generative language models with cross-lingual alignment. *arXiv preprint arXiv:2311.08089*.

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. 2025a. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. *arXiv preprint arXiv:2511.05923*.

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Libo Qin, Ruihan Chen, Baohang Li, Kui Jiang, Yaowei Wang, Ting Liu, et al. 2025b. Cai: Caption-sensitive attention intervention for mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2506.23590*.

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

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1).

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mcsqa: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. *arXiv preprint arXiv:2406.04215*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Xinyu Tang, Xiaolei Wang, Zhihao Lv, Yingqian Min, Wayne Xin Zhao, Binbin Hu, Ziqi Liu, and Zhiqiang Zhang. 2025. Unlocking general long chain-of-thought reasoning capabilities of large language models via representation engineering. *arXiv preprint arXiv:2503.11314*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025. Qwen2.5-vl.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Hongyu Wang, Jiayu Xu, Senwei Xie, Ruiping Wang, Jialin Li, Zhaojie Xie, Bin Zhang, Chuyan Xiong, and Xilin Chen. 2024. M4u: Evaluating multilingual understanding and reasoning for large multimodal models. *arXiv preprint arXiv:2405.15638*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025a. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, et al. 2025b. Polymath: Evaluating mathematical reasoning in multilingual contexts. *arXiv preprint arXiv:2504.18428*.

Junhao Xiao, Zhiyu Wu, Hao Lin, Yi Chen, Yahui Liu, Xiaoran Zhao, Zixu Wang, and Zejiang He. 2026. Not just what's there: Enabling clip to comprehend negated visual descriptions without fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 10978–10986.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zekai Ye, Qiming Li, Xiaocheng Feng, Libo Qin, Yichong Huang, Baohang Li, Kui Jiang, Yang Xiang, Zhirui Zhang, Yunfei Lu, et al. 2025. Claim: Mitigating multilingual object hallucination in large vision-language models with cross-lingual attention intervention. *arXiv preprint arXiv:2506.11073*.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. 2024a. Lingualift: An effective two-stage instruction tuning framework for low-resource language tasks. *arXiv e-prints*, pages arXiv–2412.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. 2024b. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Weixiang Zhao, Jiahe Guo, Yang Deng, Tongtong Wu, Wenxuan Zhang, Yulin Hu, Xingyu Sui, Yanyan Zhao, Wanxiang Che, Bing Qin, et al. 2025. When less language is more: Language-reasoning disentanglement makes llms better multilingual reasoners. *arXiv preprint arXiv:2505.15257*.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *Advances in Neural Information Processing Systems*, 37:15296–15319.

A Translation Quality

To evaluate the translation quality of our constructed **MMathVerse** and **MMathVista** benchmarks, we sample 500 translated queries from each benchmarks for each language and back-translate them into English using Google Translate. The back-translated English queries are then input into LVLMs to test whether their predictions align with those generated from the original English queries. High prediction consistency indicates that the translated data maintains superior benchmark quality. As shown in Table 7, these results demonstrate the reliability of our constructed multilingual dataset.

Lang.	**Zh**	**Ja**	**Es**	**Ru**	**Fr**	**De**	**Th**	**Sw**
APC	100.0	100.0	100.0	99.8	100.0	100.0	99.8	99.7

Table 7: Average Predicted Consistency (APC, %) MMathVerse and MMathVista across eight languages.

B Experimental Details

In this section, we present the experimental details, including inference settings, design of experimental prompts, fine-grained results of MathVerse & MathVista, and ablation of intervention strength.

B.1 Inference Settings

To ensure that the reasoning parameters are better aligned with reasoning tasks and to guarantee the reproducibility of results, we carefully design inference settings for each model, as shown in Table 8.

Model	**Setting**
Qwen2.5-7B-Instruct	do_sample=True, temperature=0.7, top_p=0.8, top_k=20
Qwen3-8B	do_sample=True, temperature=0.7, top_p=0.8, top_k=20, enable_thinking=False
Llama-3.1-8B- Instruct	do_sample=True
LLaVA-Onevision-7B	do_sample=True, temperature=0.6
Qwen2.5-VL-7B-Instruct	do_sample=True, temperature=0.1, top_p=0.001, repetition_penalty=1.1
InternVL3.5-VL-8B-Chat	do_sample=True, temperature=0.1, top_p=0.001, repetition_penalty=1.1

Table 8: Inference settings for each LLM and LVLM.

B.2 Design of Experimental Prompts

In this section, we provide a detailed description of each type of prompt, along with intended purposes.

Language Forcing Prompts are designed to enforce the model to generate outputs in the same language as the input. Following prior work (Wang et al., 2025b), we adopt a similar strategy, with the detailed prompt contents provided in Table 9.

Lang.	**Prompts**
En	Use English to think and answer.
Zh	使用中文进行思考和回答。
Ja	日本を使って考え、回答してください。
Es	Usa espaol para pensar y responder.
Fr	Utilisez le franais pour penser et r é pondre.
De	Verwende Deutsch, um zu denken und zu antworten.
Sw	Tumia Kiswahili kufikiri na kujibu.

Table 9: Language forcing prompts contents.

Prompts for MGSM and MSVAMP These two datasets primarily adopt free-form formats to evaluate the reasoning capabilities of LLMs. As shown in Table 10, we ask the model to first generate reasoning responses and then mark final answers using the `\\boxed{}` format.

Lang.	**Prompts**
En	Please first reason through the problem, then provide the final answer, expressed as a number using the `\\boxed{}` format.
Zh	请首先进行推理，然后给出最后的答案，用`\\boxed{}`的形式表示最后的数字。
Ja	まず推理を行ってください。その後、最的な答えを`\\boxed{}`の形式で表示してください。
Es	Por favor, primero realice el razonamiento y luego d é la respuesta final, representando el n ú mero final en forma de `\\boxed{}`.
Fr	Veuillez d'abord raisonner, puis donner la r é ponse finale sous la forme `\\boxed{}`.
De	Bitte f ü hren Sie zunchst Ihre berlegungen durch und geben Sie dann die endg ü ltige Antwort an. Verwenden Sie zur Darstellung der endg ü ltigen Zahl die Form `\\boxed{}`.
Sw	Tafadhali fanya hoja kwanza, kisha toa jibu la mwisho kwa kuandika nambari ya mwisho katika umbo la `\\boxed{}`.

Table 10: Reasoning prompts for MGSM and MSVAMP.

Prompts for MathVerse and MathVista These two benchmarks are employed to evaluate the reasoning capability of LVLMs, covering two categories of tasks: (1) multiple-choice and (2) free-form. As shown in Table 11, we design tailored reasoning prompts to guide the model's response for multiple-choice queries. The prompts for free-form are the same as the prompts for MGSM and MSVAMP benchmarks.

Lang.	Prompts
En	Please first conduct reasoning, then answer the question and put the correct option letter into \\boxed{}, e.g., \\boxed{A}, \\boxed{B}, \\boxed{C}, \\boxed{D}, at the end.
Zh	请先进行推理，然后回答问题，并在最后将正确的选项字母填入\\boxed{}中，例如：\\boxed{A}，\\boxed{B}，\\boxed{C}，\\boxed{D}.
Ja	まず推を行い、次ににに答えて、正しいオプション文字を \\boxed{}に入れます。例: \\boxed{A},\\boxed{B},\\boxed{C},\\boxed{D}, 最後に。
Es	Primero realice el razonamiento, luego responda la pregunta y coloque la letra de la opción correcta en \\boxed{}, por ejemplo, \\boxed{A}, \\boxed{B}, \\boxed{C}, \\boxed{D}, al final.
Fr	Veuillez d'abord effectuer un raisonnement, puis répondre à la question et mettre la lettre d'option correcte dans \\boxed{}, par exemple, \\boxed{A}, \\boxed{B}, \\boxed{C}, \\boxed{D}, à la fin.
De	Bitte führen Sie zuerst eine Argumentation durch, beantworten Sie dann die Frage und setzen Sie am Ende den richtigen Optionsbuchstaben in \\boxed{}, z. B. \\boxed{A}, \\boxed{B}, \\boxed{C}, \\boxed{D}.
Sw	Tafadhali kwanza elekeza hoja, kisha ujibu swali na uweke herufi ya chaguo sahihi kwenye \\boxed{}, k.m., \\boxed{A}, \\boxed{B}, \\boxed{C}, \\boxed{D}, mwishoni.

Table 11: Multi-choice reasoning prompts for MathVerse and MathVista.

B.3 Ablation of Intervention Strength

In this subsection, we conduct a detailed analysis of how the intervention strengths α_1 and α_2 influence the performance of MRRE.

We first analyze the intervention strength of the cross-lingual enhancement vectors, α_1, which aim to align non-English reasoning hidden states with English in the latent space. As shown in Table 12, the optimal performance is achieved when $\alpha_1 = 1$, while weaker or stronger interventions fail to minimize the distribution gap in the latent space, limiting the effectiveness of MRRE.

Then we analyze the intervention strength of the cross-lingual enhancement vectors, α_2, which aim to align English-like generating hidden states with the original non-English in the latent space. As shown in Table 13, the optimal performance is achieved when $\alpha_2 = 0.75$, while weaker interventions may lead to English responses rather than target non-English responses and too strong interventions may result in performance drops in reasoning tasks. Notably, α_2 functions as a trade-off strategy, effectively **balancing multilingual reasoning performance with input-output language consistency**.

α_1	0.00	0.25	0.50	0.75	**1.00**	1.25
Zh	78.4	79.3	80.5	82.3	**83.2**	82.5
Ja	65.2	67.1	68.3	71.0	**73.7**	70.8
Th	71.6	72.3	73.1	74.2	**75.2**	73.9

Table 12: Ablation study of α_1 on MGSM benchmark using Qwen2.5-7B-Instruct. Layer1 is set to be 20.

α_2	0.00	0.50	**0.75**	1.00
Zh	83.2 (26.5)	82.5 (78.7)	**81.6 (99.2)**	79.5 (99.5)
Ja	73.7 (19.8)	71.9 (69.9)	**71.2 (99.1)**	70.0 (99.6)
Th	75.2 (7.9)	75.0 (54.1)	**74.8 (99.2)**	72.1 (99.4)

Table 13: Ablation study of α_2 on MGSM benchmark using Qwen2.5-7B-Instruct. The first numbers denote accuracies (%) and the second numbers denote Language Consistency (LC, %). Layer1 is set to be 20, α_1 is set to be 1.0, and Layer2 is set to be 23.

B.4 Fine-grained Results of MathVerse and MathVista Benchmarks

To further elucidate how MRRE improves LVLM reasoning, we perform a fine-grained analysis across eight languages, five visual task families (*text-dominant, text-lite, vision-integrated, vision-dominant, and vision-only*), and two domains (*general* and *math*). We disaggregate performance and consistency metrics by language, task category, and domain to identify where MRRE yields the largest gains; results are reported in Tables 14.

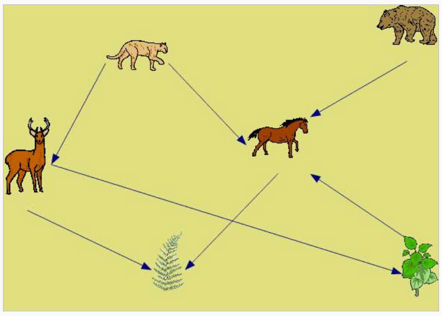
C Fine-grained Case Studies

To better illustrate the improvements in reasoning capabilities and input-output language consistency achieved by MRRE, we provide additional comprehensive case studies.

Model	*MMathVerse*						*MathVista*		
	T-D	**T-L**	**V-I**	**V-D**	**V-O**	**Overall**	**General**	**Math**	**Overall**
Language-Zh									
Qwen2.5-VL-7B	44.4	39.5	35.5	36.0	36.5	38.4	59.3	57.0	58.1
+ Language forcing	41.8	35.3	33.4	34.6	38.7	36.8	**60.8**	59.8	60.0
+ MRRE	**54.2**	**47.0**	**43.8**	**42.9**	**41.9**	**45.9**	60.7	**63.7**	**62.3**
Language-Ja									
Qwen2.5-VL-7B	42.8	35.3	33.5	34.5	35.0	36.4	55.0	49.4	52.0
+ Language forcing	44.2	39.2	36.8	36.5	39.6	37.7	54.8	51.5	53.0
+ MRRE	**51.6**	**46.2**	**42.5**	**42.4**	**40.5**	**44.6**	**59.3**	**62.6**	**61.1**
Language-Es									
Qwen2.5-VL-7B	49.5	43.5	40.9	38.1	40.4	42.5	63.0	61.5	62.2
+ Language forcing	51.8	43.4	40.5	42.3	40.1	43.6	62.2	60.4	61.2
+ MRRE	**55.5**	**49.5**	**46.4**	**45.1**	**40.7**	**47.5**	**64.1**	**64.3**	**64.2**
Language-Ru									
Qwen2.5-VL-7B	45.1	39.0	35.8	35.8	37.6	38.6	60.4	55.7	57.9
+ Language forcing	46.7	41.2	39.6	38.1	37.4	40.6	61.3	55.7	58.3
+ MRRE	**52.0**	**45.3**	**41.8**	**42.3**	**42.4**	**44.7**	**62.8**	**62.8**	**62.8**
Language-Fr									
Qwen2.5-VL-7B	43.0	38.6	35.8	36.8	36.4	38.1	61.5	58.0	59.6
+ Language forcing	46.6	40.0	36.9	36.5	37.9	39.6	61.1	58.0	59.4
+ MRRE	**51.9**	**47.5**	**44.7**	**42.0**	**42.9**	**45.8**	**62.2**	**60.2**	**61.2**
Language-De									
Qwen2.5-VL-7B	42.5	38.2	32.5	33.4	36.0	36.5	58.9	57.6	58.2
+ Language forcing	50.5	39.0	35.3	36.9	38.2	40.0	56.1	57.6	56.9
+ MRRE	**52.9**	**47.0**	**42.9**	**43.8**	**41.6**	**45.7**	**63.5**	**64.6**	**64.1**
Language-Th									
Qwen2.5-VL-7B	36.8	33.1	31.2	31,2	27.5	32.0	52.0	50.4	51.1
+ Language forcing	50.3	43.8	41.8	40.6	41.0	43.5	55.9	51.9	53.7
+ MRRE	**56.7**	**51.0**	**43.8**	**43.8**	**46.9**	**48.4**	**58.7**	**55.8**	**57.3**
Language-Sw									
Qwen2.5-VL-7B	15.1	13.8	11.8	14.0	27.3	16.4	17.6	15.9	16.7
+ Language forcing	20.1	16.0	15.0	15.0	22.8	17.8	8.7	11.3	10.1
+ MRRE	**34.4**	**32.5**	**31.6**	**32.4**	**38.2**	**33.8**	**20.4**	**35.6**	**28.6**

Table 14: Fine-grained Results on Qwen2.5-VL-7B-Instruct across eight languages.

Input Image



Input Query w/ English

In the diagram of the food web shown what will most directly be affected by the loss of the trees?
Choices:
(A) horses
(B) cats
(C) nothing
(D) bears

Qwen2.5-VL-7B w/ Es Query

Para responder esta pregunta, debemos analizar cómo los árboles están conectados en la red trófica mostrada.\n\n1. Los árboles son una fuente de alimento para los animales herbívoros.\n2. En este diagrama, los animales herbívoros son los ciervos (que comen hojas de árbol).\n3. Los caballos también comen hojas de árbol, pero no están directamente conectados a los árboles como lo están los ciervos.\n4. Los gatos y los osos no tienen conexión directa con los árboles en este diagrama.\n\nPor lo tanto, la pérdida de árboles tendrá un impacto más directo en los ciervos, ya que son los únicos animales que dependen directamente de los árboles como fuente de alimentación.\n\nLa respuesta correcta es:\n\n**\\boxed{B}**.

Qwen2.5-VL-7B+MRRE w/ Es Query

Para determinar qué organismo se verá más directamente afectado por la pérdida de árboles en el diagrama de la red alimentaria dada, necesitamos analizar las flechas que van desde los árboles hacia otros organismos. Estas flechas indican la dirección del flujo de energía y quién consume a quién.\n\n1. Los árboles tienen flechas que apuntan hacia ellos desde los ciervos y las plantas, lo que indica que estos organismos consumen los árboles.\n2. Hay una flecha que va desde los árboles hacia las plantas, lo que sugiere que las plantas también consumen los árboles.\n3. Las plantas tienen flechas que apuntan hacia los caballos, lo que indica que los caballos consumen las plantas.\n4. Los caballos tienen una flecha que apunta hacia los osos, lo que indica que los osos consumen a los caballos.\n5. Los gatos tienen una flecha que apunta hacia los ciervos, lo que indica que los ciervos consumen a los gatos.\n6. Los osos no tienen ninguna flecha que apunte hacia ellos desde los árboles, por lo que no se ven directamente afectados por la pérdida de árboles.\n\nA partir de este análisis, está claro que las plantas se ven directamente afectadas por la pérdida de árboles porque hay una flecha que va desde los árboles hacia las plantas. Sin embargo, dado que la pregunta pide el organismo que se verá más directamente afectado, y las plantas son consumidas por los caballos, los caballos se verán afectados indirectamente a través de las plantas.\n\nPor lo tanto, la respuesta es:**\\boxed{A}**.

Qwen2.5-VL-7B w/ Ru Query

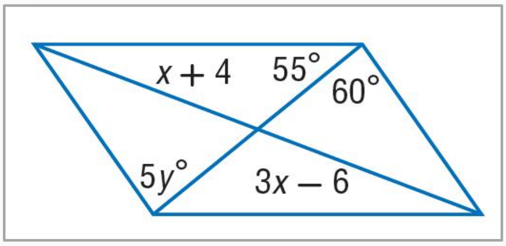
Для ответа на этот вопрос нужно внимательно рассмотреть схему пищевой цепи:\n\n1. Деревья являются основным источником питания для травоядных животных (лошадь и олень).\n2. Лошадь и олень, в свою очередь, становятся пищей для хищников (кошка).\n\nТаким образом, если исчезнут деревья, то это приведет к недостатку пищи для лошадей и оленей, а следовательно, и для кошек. Медведи не указаны в этой пищевой цепи как потребители растительной пищи.\n\nОтвет: **\\boxed{C}**.

Qwen2.5-VL-7B+MRRE w/ Ru Query

Чтобы определить ответ на этот вопрос, нам нужно проанализировать пищевую сеть, изображённую на рисунке, и понять, какой организм напрямую пострадает от исчезновения деревьев (продуцентов в основании пищевой сети).\n\n1. Определение продуцентов: Растения в нижней части диаграммы представляют собой основных продуцентов, таких как деревья и кустарники.\n\n2. Определение прямых потребителей: Организмы, которые питаются этими продуцентами, — это травоядные, в частности лошади и, возможно, другие мелкие животные, не показанные на схеме, но подразумеваемые из контекста.\n\n3. Определение хищников: Хищники на диаграмме включают кошку и медведя, которые охотятся на травоядных, таких как лошадь.\nУчитывая эту информацию, если деревья исчезнут, прямое воздействие будет на травоядных, которые зависят от них в питании. В данной диаграмме таким травоядным является лошадь.\n\nТаким образом, правильный ответ:**\\boxed{A}.**

Figure 7: Case study of *General* problem on Qwen2.5-VL-7B-Instruct.

Input Image



Input Query w/ English

Find x so that the quadrilateral is a parallelogram.
Choices:
A:3
B:5
C:12
D:18

Qwen2.5-VL-7B w/ Chinese Query

为了判断这个图形是否是一个平行四边形，我们需要利用平行四边形的性质和角度关系。
首先，我们知道平行四边形对角相等且邻角互补（即两个相邻的角度之和等于180度）。
观察图中的角度信息：一个角是 \\(60^\\circ\\)，另一个与它相邻的角是 \\(55^\\circ + (x+4)\\) 或者 \\((3x - 6)\\)，因为它们是平行线上的同旁内角，所以这两个角应该互补，即 \\(60^\\circ + [55^\\circ + (x+4)] = 180^\\circ\\) 或者 \\(60^\\circ + [(3x - 6)] = 180^\\circ\\)。
我们来解第一个方程：
\[60^\circ + 55^\circ + (x + 4) = 180^\circ\]
\[119^\circ + x + 4 = 180^\circ\]
\[x + 123 = 180^\circ\]
\[x = 180^\circ - 123^\circ\]
\[x = 57^\circ - 123^\circ\]
显然，上面的计算有误，因为我们没有正确地应用补角的概念。实际上，我们应该直接使用第二个方程来解决这个问题，因为它更直接反映了一个邻角的关系。
现在考虑第二个方程：
\[60^\circ + (3x - 6) = 180^\circ\]
\[60^\circ + 3x - 6 = 180^\circ\]
\[3x + 54 = 180^\circ\]
\[3x = 180^\circ - 54^\circ\]
\[3x = 126^\circ\]
\[x = 42^\circ\]
但是，根据题目给出的答案选项来看，我们的目标是找到一个合适的 x 值使得图形成为平行四边形。因此，我们需要重新审视原始条件并确保选择正确的答案。考虑到题目提供的选项，我们可以尝试验证每个选项是否满足条件。但基于上述分析，最接近的答案应该是通过验证选项得出的。
最终答案应为 \(\boxed{D}\)。

Qwen2.5-VL-7B+MRRE w/ Chinese Query

为了确定使给定四边形成为平行四边形的 \(x \) 的值，我们需要利用平行四边形的性质。具体来说，在平行四边形中，\textbf{对边长度相等}。

已知边长为 \(x + 4 \)、\(5y \)、\(3x - 6 \)，以及另一条未标记但可根据问题设置推断的边，我们需要根据对边相等的性质来建立方程。

从图表中，我们有：
- 一条边是 \(x + 4 \)。
- 其对边是 \(3x - 6 \)。

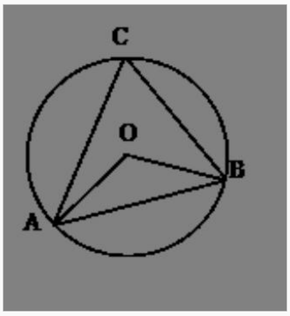
将这两个表达式设为相等，得到：
\[x + 4 = 3x - 6 \]

现在，求解 \(x \)：
1. 两边同时减去 \(x \)：
\[4 = 2x - 6 \]
2. 两边同时加上 6：
\[10 = 2x \]
3. 两边同时除以 2：
\[x = 5 \]

因此，使该四边形成为平行四边形的 \(x \) 的值是 **\(\boxed{B}\)**。

Figure 8: Case study of *Text Lite* problem on Qwen2.5-VL-7B-Instruct.

Input Image



Input Query w/ English

As shown in the figure, triangle ABC is the inscribed triangle of circle O, angle OAB = 35.0, then the degree of angle ACB is ()

Choices:

A:35 °

B:55 °

C:60 °

D:70 °

Qwen2.5-VL-7B w/ Zh Query

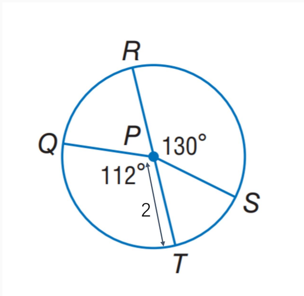
首先，我们知道在圆中，如果一个角的顶点在圆心上，则这个角所对的弧等于该角的一半。\n\n题目给出的是角OAB = 35.0 °，这是一个圆周角，它所对的弧是AB。根据圆周角定理，圆周角等于它所对的圆心角的一半，所以角AOB = 2 * 角OAB = 2 * 35.0 ° = 70.0 °。\n\n接下来，我们需要找出角ACB的度数。由于角ACB是一个圆周角，它所对的弧也是AB。因此，角ACB也等于角AOB的一半，即角ACB = 1/2 * 角AOB = 1/2 * 70.0 ° = 35.0 °。\n\n所以，正确答案是 A: 35 °。**\n\n\\boxed{A}**。

Qwen2.5-VL-7B+MRRE w/ Zh Query

为了解决这个问题，我们需要利用圆和三角形的性质。\n\n1. **确定已知条件：**\n - 三角形 \(\triangle ABC \) 内接于圆 \(O \)。\n - \(\angle OAB = 35^\circ \)。\n\n2. **理解圆心角与内接角的关系：**\n - 内接角的度数等于它所对的圆心角度数的一半。\n - 在本题中，\(\angle ACB \) 是一个内接角，它所对的弧与圆心角 \(\angle AOB \) 相同。\n\n3. **求 \(\angle AOB \) 的度数：**\n - 因为 \(\angle OAB \) 属于三角形 \(\triangle OAB \)，且已知 \(OA = OB \)（都是圆的半径），所以 \(\triangle OAB \) 是等腰三角形，且 \(\angle OAB = \angle OBA = 35^\circ \)。\n - 因此，三角形 \(\triangle OAB \) 的内角和应为 \(180^\circ \)：\n \[\n \angle AOB + \angle OAB + \angle OBA = 180^\circ\n \]\n 代入已知数值：\n \[\n \angle AOB + 35^\circ + 35^\circ = 180^\circ\n \]\n 化简：\n \[\n \angle AOB + 70^\circ = 180^\circ\n \]\n 解得：\n \[\n \angle AOB = 110^\circ\n \]\n\n4. **求 \(\angle ACB \) 的度数：**\n - 因为 \(\angle ACB \) 是一个内接角，它所对的弧与圆心角 \(\angle AOB \) 相同，所以 \(\angle ACB \) 的度数是 \(\angle AOB \) 的一半：\n \[\n \angle ACB = \frac{1}{2} \times \angle AOB\n \]\n 代入 \(\angle AOB \) 的数值：\n \[\n \angle ACB = \frac{1}{2} \times 110^\circ = 55^\circ\n \]\n\n因此，正确答案是：**\n\n\[\n\boxed{B}\n\]**

Figure 9: Case study of *Text Dominant* problem on Qwen2.5-VL-7B-Instruct.

Input Image



Input Query w/ English

In \\odot P, find the length of \\widehat R S. Round to the nearest hundredth.
Choices:
A:2.27
B:3.91
C:4.54
D:7.82

Qwen2.5-VL-7B w/ Ja Query

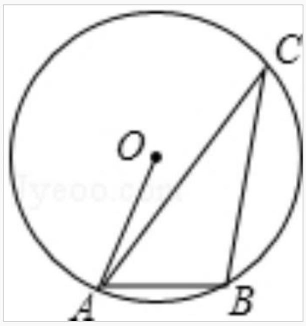
この問題では、円周角の定理と弧の長さの公式を使用します。\n\n1. **円周角の定理**: 円周角は中心角の半分になります。つまり、\\(\\angle RPS = \\frac{1}{2} \\times 130^\\circ = 65^\\circ\\) です。\n\n2. **弧の長さの公式**: 弧の長さ \\(L\\) は、次の式で計算できます:\n \\[\n L = r \\theta\n \\]\n ここで, \\(r\\) は半径, \\(\\theta\\) は中心角（ラジアン）です。\n\n - 半径 \\(r = 2\\)\n - 中心角 \\(\\theta = 65^\\circ\\) をラジアンに変換します。\\(1^\\circ = \\frac{\\pi}{180}\\) ラジアンなので,\n \\[\n \\theta = 65^\\circ \\times \\frac{\\pi}{180} = \\frac{65\\pi}{180} = \\frac{13\\pi}{36}\n \\]\n\n - 弧の長さ \\(L\\) を計算します:\n \\[\n L = 2 \\times \\frac{13\\pi}{36} = \\frac{26\\pi}{36} = \\frac{13\\pi}{18}\n \\]\n\n 約数を計算すると:\n \\[\n \\frac{13\\pi}{18} \\approx \\frac{13 \\times 3.1416}{18} \\approx \\frac{40.8708}{18} \\approx 2.2706\n \\]\n\n 小数点第1位を四捨五入すると約 2.3 ですが、最も近い選択肢は A:2.27 です。\n\nしたがって、答えは \\boxed{A} です。

Qwen2.5-VL-7B+MRRE w/ Ja Query

円 \(\odot P \) でアーク RS の長さを見つけるには、中央の角度とそれが切断するアークの関係を使用する必要があります。弧の測定は、その対応する中央角に比例します。アークの長さの公式は:nn\[\text{アーク長さ} = \frac{\theta}{360^\circ} \times 2\pi r \]nnwhere:n- \(\theta \) は中央角度の測定値,n- \(r \) は円の半径です.ダイアグラムから、中心角 \(\angle RPS \) は \(130^\circ \) として与えられていることがわかります。半径 \(r \) は 2 単位として与えられます。これらの値を公式に置き換える:nn\[\text{アーク長さ} = \frac{130^\circ}{360^\circ} \times 2\pi \times 2 \]nn表現を簡単にします:nn\[\text{アーク長さ} = \frac{130}{360} \times 4\pi \]n\[\text{アーク長さ} = \frac{13}{36} \times 4\pi \]n\[\text{アーク長さ} = \frac{52\pi}{36} \]n\[\text{アーク長さ} = \frac{13\pi}{9} \]nn次に、(\pi \approx 3.14159 \):nn\[\text{アーク長さ} = \frac{13 \times 3.14159}{9} \]n\[\text{アーク長さ} \approx \frac{40.84477}{9} \]n\[\text{アーク長さ} \approx 4.54 \]nn最寄りの100番目にめること (必要に応じて最寄りの10番目にめることが必要であるため。

Figure 10: Case study of *Vision Dominant* problem on Qwen2.5-VL-7B-Instruct.

Input Image



Input Query w/ English

As shown in the figure, angle ACB = 25.0, then the degree of angle BAO is ()
Choices:
A:50 °
B:55 °
C:60 °
D:65 °

Qwen2.5-VL-7B w/ Sw Query

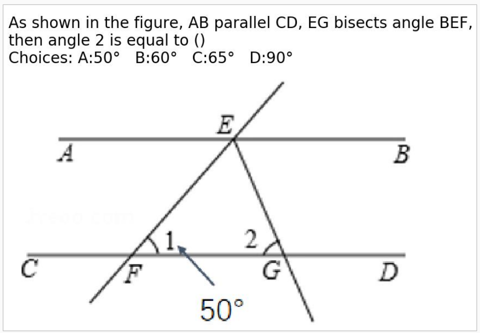
Kuna njia za mbili kuhusu hii:\n\n1. **Naangalia kutumika namba za maelezo**: Angaliwezi kuongeza anasema kwamba angle ACB = 25 ° .\n\n2. **Kutumika namba za maelezo**:\n - Angaliwezi kuongeza anasema kwamba basi kiwango cha angle BAO ni angle AOB.\n - Kwa sababu angle AOB ni angle inscribed (kujulikana kama angle subtended by the arc AB) ambayo ni equal to twice the angle at the circumference that subtends the same arc.\n \n3. **Kutumika namba za maelezo**:\n - Angaliwezi kuongeza anasema kwamba basi kiwango cha angle BAO ni 2 * angle ACB.\n - Tofauti na angle ACB = 25 ° , basi kiwango cha angle BAO ni 2 * 25 ° = 50 ° .\n\nKwa hivyo, basi kiwango cha angle BAO ni 50 ° .\n\nJinsi yote, angaliwezi kuongeza anasema kwamba:\n\n\\boxed{A}

Qwen2.5-VL-7B+MRRE w/ Sw Query

Ili kutatua tatizo hili, tunahitaji kutumia mali za duara na pembe katika pembetatu.\n\nGiven:\n- \(\angle ACB = 25^\circ \)\n- Tunahitaji kupata kipimo cha \(\angle BAO \)\n\nUfafanuzi wa hatua kwa hatua:\n\n1. Tambua aina ya pembetatu: Kwa kuwa \(O \) ni kituo cha duara, \(OA \) na \(OB \) ni nusu kipenyo cha duara. Hivyo basi, \(\triangle AOB \) ni pembetatu ya isosceles yenye \(OA = OB \).\n\n2. Jumla ya pembe katika pembetatu: Jumla ya pembe za ndani katika pembetatu yoyote ni \(180^\circ \). Katika \(\triangle AOB \), pembe kwenye \(O \) ni sawa kwa sababu ni pembetatu ya isosceles. Tuita hizi pembe \(\angle OAB \) na \(\angle OBA \). Kwa hiyo, tunayo:\n \[\n \angle OAB + \angle OBA + \angle AOB = 180^\circ\n \]\n\n3. Uhusiano kati ya pembe ya katikati na pembe kwenye duara: Pembe ya katikati \(\angle AOB \) inakaa kwenye upinde ule ule kama pembe kwenye duara \(\angle ACB \). Uhusiano ni kwamba pembe ya katikati ni mara mbili ya pembe ya duara. Kwa hivyo:\n \[\n \angle AOB = 2 \times \angle ACB\n \]\n Tukibadilisha thamani iliyotolewa:\n \[\n \angle AOB = 2 \times 25^\circ = 50^\circ\n \]\n\n4. Jumla ya pembe katika \(\triangle AOB \): Sasa tunajua moja ya pembe katika \(\triangle AOB \):\n \[\n \angle AOB = 50^\circ\n \]\n Kwa kuwa \(\triangle AOB \) ni ya isosceles, basi:\n \[\n \angle OAB = \angle OBA\n \]\n Tuita kila moja ya hizi pembe \(x \). Kisha:\n \[\n x + x + 50^\circ = 180^\circ\n \]\n Tukirahisisha:\n \[\n 2x + 50^\circ = 180^\circ\n \]\n Tukitatua kwa \(x \):\n \[\n 2x = 130^\circ\n \]\n \[\n x = 65^\circ\n \]\n\nKwa hivyo, kipimo cha \(\angle BAO \) ni \(65^\circ \).\n\nJibu sahihi ni:\n\[\n\boxed{D}\n\]

Figure 11: Case study of *Vision Intensive* problem on Qwen2.5-VL-7B-Instruct.

Input Image



No Input Query

Qwen2.5-VL-7B w/ Fr Query

Pour résoudre ce problème, nous devons utiliser les propriétés des angles formés par deux droites parallèles coupées par une transversale.\n\n1. **Identifiez les angles correspondants :**\n - Les angles 1 et 50 ˚ sont des angles alternes-internes car ils sont situés de part et d'autre de la ligne transversale EG et sur le même côté du plan par rapport aux droites AB et CD.\n \n2. **Utilisez la propriété des angles alternes-internes :**\n - Si deux droites sont parallèles, alors les angles alternes-internes sont égaux.\n - Donc, angle 1 = 50 ˚ .\n\n3. **Angle bisecteur :**\n - EG est un angle bisecteur qui divise l'angle BEF en deux angles égaux.\n - Donc, angle BEG = angle FEG = (angle BEF) / 2.\n\n4. **Calculer l'angle BFG :**\n - Angle BFG est l'angle supplémentaire de l'angle 1 (car ils forment un angle droit avec l'angle 50 ˚).\n - Donc, angle BFG = 180 ˚ - 50 ˚ = 130 ˚ .\n\n5. **Angle 2 :**\n - Angle 2 est l'angle interne adjacent à l'angle BFG.\n - Donc, angle 2 = 180 ˚ - angle BFG = 180 ˚ - 130 ˚ = 50 ˚ .\n\nDonc, l'angle 2 est égal à 50 ˚ .\n\nLa réponse est: **\\boxed{A}.**

Qwen2.5-VL-7B+MRRE w/ Fr Query

Pour résoudre le problème, nous devons utiliser les propriétés des droites parallèles et des bissectrices d'angles.\n\n1. **Identifier lesinformations données :**\n - \(AB \parallel CD\)\n - \(EG\) est la bissectrice de \(\angle BEF \)\n - \(\angle 1 = 50^\circ \)\n\n2.**Comprendre la relation entre les angles :**\n - Puisque \(AB \parallel CD\), \(\angle BFE \) est un angle extérieur pour \(\triangleEFG \) et il est égal à la somme des deux angles intérieurs non adjacents.\n - Le théorème de la bissectrice stipule que la bissectriced'un angle divise le côté opposé en segments proportionnels aux côtés adjacents. Cependant, dans ce contexte, nous pouvons utiliserle fait que la bissectrice divise l'angle en deux parties égales.\n\n3. **Calculer \(\angle BEF \) :**\n - Comme \(\angle 1 = 50^\circ\) et que \(EG\) est la bissectrice de \(\angle BEF \), posons \(\angle BEG = \angle GEF = x \).\n - Donc, \(\angle BEF = 2x \).\n\n4.**Utiliser la propriété des angles correspondants :**\n - Comme \(AB \parallel CD\), \(\angle BFE = \angle 1 = 50^\circ \).\n - Donc,\(2x = 50^\circ\).\n - En résolvant pour \(x\), on obtient \(x = 25^\circ\).\n\n5. **Trouver \(\angle 2 \) :**\n - \(\angle 2 \) est lemême que \(\angle GEF \) car ce sont des angles correspondants formés par la sécante \(EG\) coupant les droites parallèles \(AB\) et\(CD\).\n - Donc, \(\angle 2 = 25^\circ \).\n\nCependant, les options fournies ne correspondent pas directement à notre valeur calculée.Réévaluons le problème avec l'approche correcte :\n\nPuisque \(\angle BFE = 50^\circ \) et que \(EG\) est la bissectrice de \(\angleBEF \), le calcul correct devrait être :\n\[\n\angle BEF = 180^\circ - 50^\circ = 130^\circ\n\]\nMais comme \(\angle 2 \) est l'angle del'autre côté du segment \(EG\), il doit être égal à la moitié de l'angle restant après avoir soustrait l'angle bisecté de 180 degrés(puisque la somme des angles sur une ligne droite est 180 degrés) :\n\[\n\angle 2 = \frac{180^\circ - 50^\circ}{2} = \frac{130^\circ}{2}= 65^\circ\n\]\n\nDonc, la réponse correcte est :**\n\\boxed{C}\n.**

Figure 12: Case study of *Vision Only* problem on Qwen2.5-VL-7B-Instruct.

D Language Consistency (LC)

To evaluate the model's proficiency in adhering to the specified target language, we define the *Language Consistency* (LC) metric as the proportion of model responses successfully generated in the target language. Given the complexity of multilingual reasoning, where responses often contain a mixture of languages or technical symbols, we implement a character-weighted majority voting mechanism supported by FastText to ensure a noise-free evaluation.

As detailed in Algorithm 1, our method first undergoes a preprocessing stage to strip LaTeX environments, mathematical commands, and URLs, preventing these elements from biasing the detector. A response is categorized as "consistent" only if the target language occupies more than 80% of the total character count based on sentence-level decomposition.

Algorithm 1 Calculation of Language Consistency (LC)

Input : Model responses $\mathcal{D}$ for target language L
Output : LC score $(\%)$

```
total_correct ← 0
foreach response R ∈ D do
    // Stage 1: Preprocessing
    R_clean ← Remove LaTeX math, commands, and URLs from R
    Clean extra symbols while preserving target scripts
    // Stage 2: Decomposition & Detection
    Split R_clean into sentence-level chunks {C_1, C_2, ..., C_n}
    len_valid ← 0, len_total ← 0
    foreach chunk C_i do
        Predict language P_i and confidence Prob_i using FastText
        if P_i == L and Prob_i ≥ 0.8 then
            len_valid ← len_valid + length(C_i)
        len_total ← len_total + length(C_i)
    // Stage 3: Sample Decision
    Ratio ← len_valid / len_total
    if Ratio > 0.8 then
        total_correct ← total_correct + 1
return (total_correct / |D|) × 100%
```

E Usage of LLMs

We used GPT-4o (Hurst et al., 2024) to assist in language refinement and readability improvement of the manuscript. All ideas, experiments, analyses, and conclusions are developed and verified by the authors.