ICLR
International Conference On Learning Representations

Beyond English-Centric Training: How Reinforcement Learning Improves Cross-Lingual Reasoning

Shulin Huang, Yiran Ding, Junshu Pan, Yue Zhang*

huangshulin@westlake.edu.cn

Zhejiang University, Westlake University

Motivation

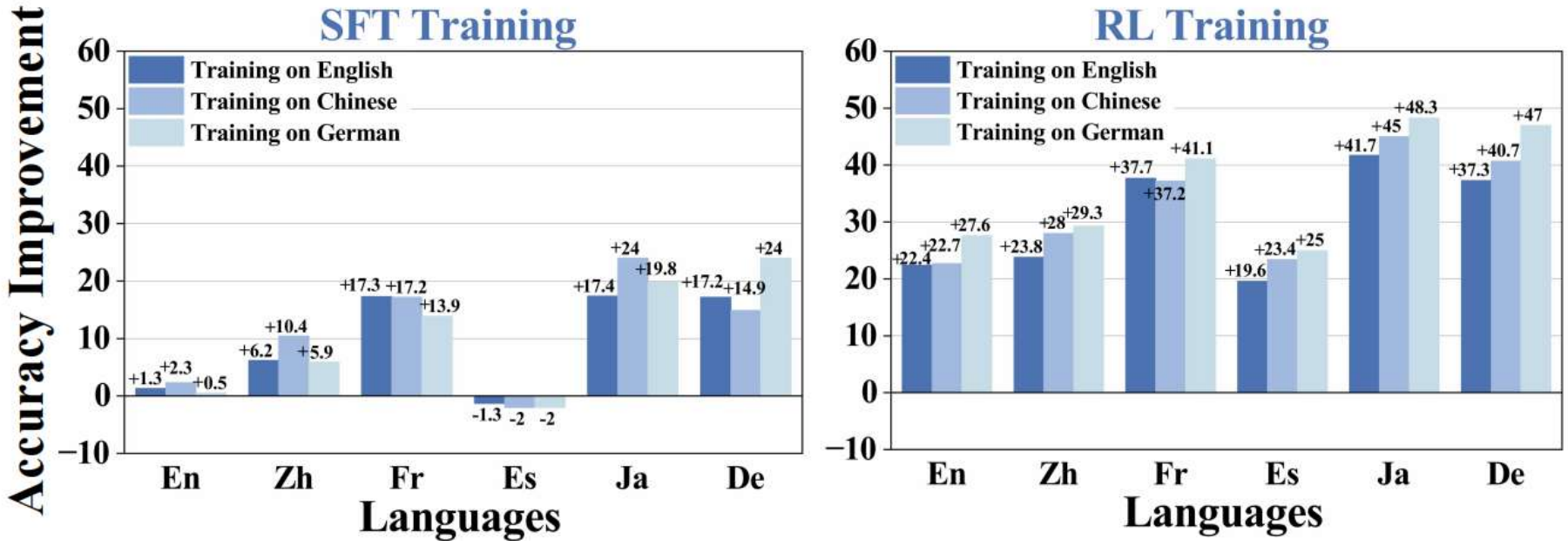
- While RL improves complex reasoning, its impact on cross-lingual generalization compared to Supervised Fine-Tuning (SFT) remains unexplored.
- Pre-training is predominantly English-centric, leading to significant performance gaps in other languages.
- Multilingual reasoning is critical for equitable global AI benefits.

Research Questions (RQs)

- **RQ1:**Does RL exhibit stronger cross-lingual reasoning generalization than SFT?
- **RQ2:**Can non-English data training perform better than English-centric training in an RL framework?
- **RQ3:**What are the underlying mechanics (linguistic consistency, sampling, semantic shift) behind RL's performance?

Key Finding 1: RL Superiority over SFT

- **Performance Gap:** RL consistently and significantly outperforms SFT across all languages.
- **Cross-Lingual Transfer:** RL advantage is most prominent in cross-lingual transfer, suggesting it learns more robust, language-agnostic strategies.



Key Finding 2: The Non-English Advantage

- **Counter-intuitive Phenomenon:** RL training on non-English data (German/Chinese) yields better overall performance than training on English data.
- **SFT Contrast:** This phenomenon is NOT observed in SFT, where language variation has minimal impact.

Table 1: Performance of base, SFT, and RL models on MGSM. "Base" denotes Qwen2.5-3B-Base. "SFT (zh)" and "RL (zh)" indicate tuning on Chinese data. We report accuracy on 10 linguistic settings; Δ (RL–SFT) denotes the performance gap. Each value is averaged over six runs. "Avg" and "Gen" refer to the mean accuracy and generalization score, respectively.

Models	En	Zh	De	Es	Fr	Ja	Ru	Th	Sw	Bn	Avg	Gen
Base	63.4	48.3	33.5	57.7	38.9	19.5	30.3	17.6	7.3	1.2	31.8	0.0
SFT (En)	64.7	54.5	50.7	56.4	56.2	36.9	55.5	44.1	6.9	26.2	45.2	18.1
RL (En)	85.8	72.1	70.8	77.3	76.6	61.2	64.9	61.0	9.5	47.5	62.7	49.1
Δ (RL-SFT)	+21.1	+17.6	+20.1	+20.9	+20.4	+24.3	+9.4	+16.9	+2.6	+21.3	+17.5	+30.9
SFT (Zh)	65.7	58.7	48.4	55.7	56.1	43.5	56.6	45.8	7.5	30.5	46.9	20.4
RL (Zh)	86.1	76.3	74.2	81.1	76.1	64.5	78.1	64.9	10.3	48.3	66.0	52.6
Δ (RL-SFT)	+20.4	+17.6	**+25.8**	**+25.4**	+20.0	+21.0	+21.5	+19.1	+2.8	+17.8	+19.1	+32.3
SFT (De)	63.9	54.2	57.5	55.7	52.8	39.3	55.1	47.6	8.4	28.8	46.3	19.3
RL (De)	91.0	77.6	80.5	82.7	80.0	67.8	81.3	75.3	15.9	63.3	71.5	60.4
Δ (RL-SFT)	**+27.1**	**+23.4**	+23.0	+27.0	**+27.2**	**+28.5**	**+26.2**	**+27.7**	**+7.5**	**+34.5**	**+25.2**	**+41.2**

Robustness Across Scales and Tasks

- **Model Scaling:** Trends hold for larger models.

- **Task Generalization:** Non-English RL also excels in out-of-distribution tasks like commonsense (MMLU-ProX-Lite) and scientific reasoning (MGPQA-D).

Table 3: Performance comparison on MMLU-ProX-Lite and MGPQA-D. "Avg" denotes the average score across languages (En/Zh/De), and "Gen" represents the generalization score.

Model	MMLU-ProX-Lite					MGPQA-D				
	En	Zh	De	Avg	Gen	En	Zh	De	Avg	Gen
Base	9.2	2.4	3.6	5.0	0.0	21.1	20.0	20.7	20.6	0.0
SFT(En)	28.9	13.0	24.1	22.0	17.9	12.5	5.1	11.7	9.8	-13.7
RL(En)	40.6	31.6	25.6	32.6	29.1	30.0	23.5	25.2	26.2	7.0
Δ (RL-SFT)	+11.6	+18.6	+1.6	+10.6	+11.2	+17.4	**+18.4**	+13.5	+16.4	+20.7
SFT(Zh)	24.4	11.6	21.3	20.3	7.4	20.7	18.1	14.2	17.7	-3.7
RL(Zh)	40.8	35.0	34.4	36.7	33.4	25.0	27.3	28.3	26.9	7.8
Δ (RL-SFT)	+16.3	**+23.4**	+13.1	+16.4	+19.8	+4.3	+9.2	+14.1	+9.2	+11.5
SFT(De)	15.8	7.3	15.0	12.7	8.0	7.2	12.8	8.8	9.6	-13.9
RL(De)	39.9	27.2	35.5	34.2	30.8	26.2	27.2	25.3	26.2	7.1
Δ (RL-SFT)	**+24.1**	+20.0	**+20.5**	**+21.5**	**+22.8**	**+19.0**	+14.4	**+16.6**	**+16.7**	**+21.0**

Mechanic 1: Language Inconsistency

- **Observation:** RL-De models do not strictly adhere to German; they spontaneously use mixed languages (like English) for reasoning.

- **Hypothesis:** This linguistic flexibility enables the model to leverage more powerful internal reasoning modules.

Constraint vs. Performance

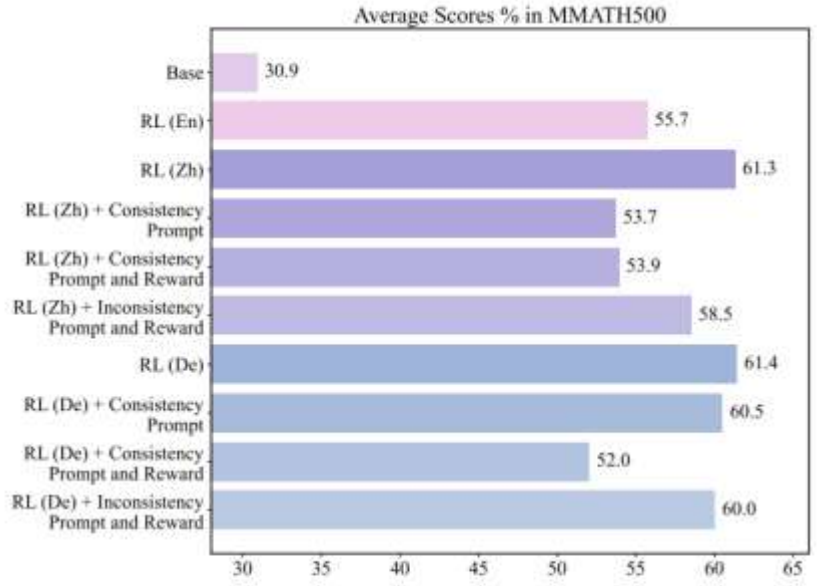


Figure 2: Scores on MMath500. The chart compares the average accuracy of different models. "RL (Zh)" indicates training on Chinese data.

- **The Penalty of Consistency:** Enforcing language consistency via prompts or rewards significantly degrades performance.
- **Consistency Rates:** RL(De) drops from 61.4% to 52.0% when forced to stay in-language.
- Language inconsistency is a potential source of cross-lingual generalization.

Mechanic 2: Role of Sampling

- **Beyond Imitation:** While SFT memorizes expert trajectories, RL's online optimization explores more diverse and effective reasoning paths.
- **Uncertainty:** Higher perplexity in German questions potentially prompts the model to explore better cross-lingual paths.

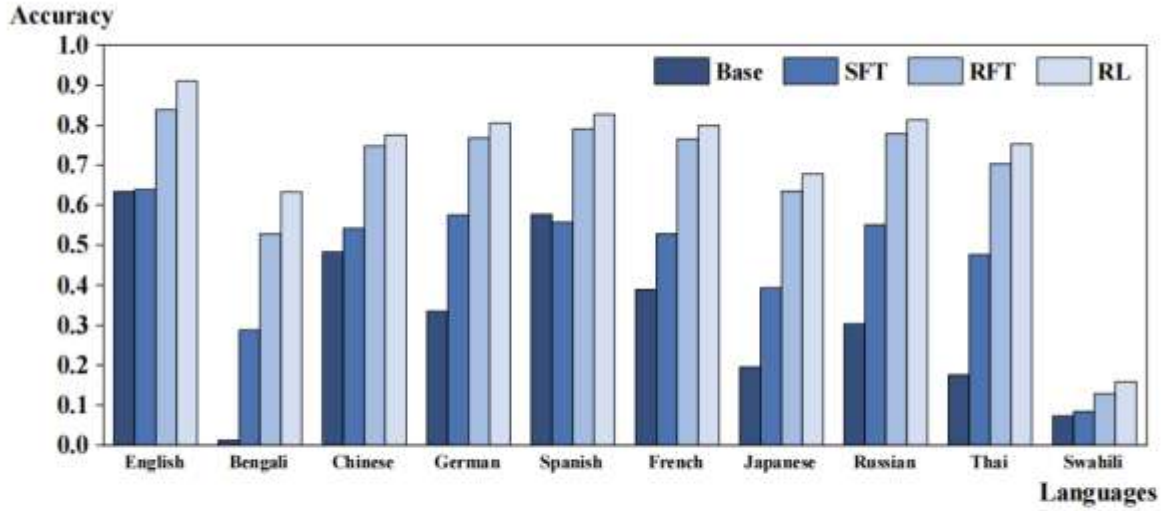


Figure 3: Model performance comparisons among the Base, SFT, RFT, and RL models on MGSM. We use German data in LUFFY in SFT, RL, RFT for training.

Mechanic 3: Semantic Feature Shift

- **The Stability factor:** RL-De exhibits a concentrated distribution, indicating minimal deviation from base pre-trained structures.
- **Insight:** Preserving pre-trained multilingual structures is crucial for robust transfer.

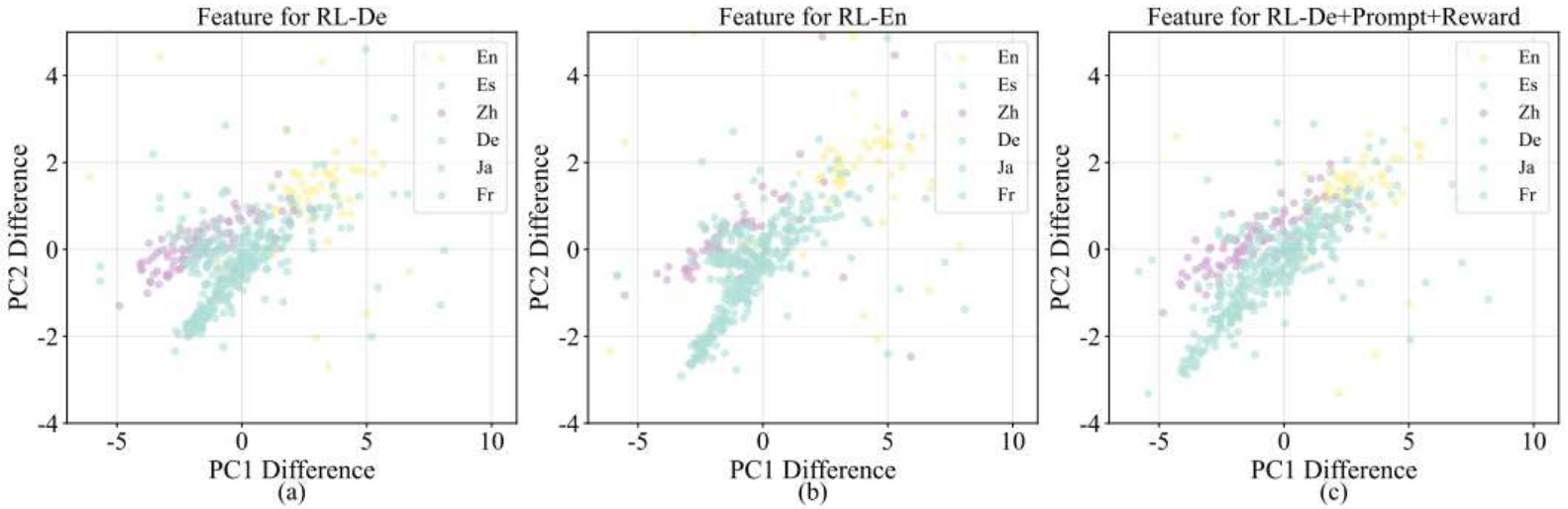


Figure 4: Feature of LLM's hidden state of last layer, in training (dataset LUFFY) configuration of (a)RL-De, (b)RL-De+Prompt, and (c))RL-De+Prompt+Reward. "+Prompt" adds language control prompts, and "+Reward" adds a language consistency reward.

Conclusion

- RL achieves superior cross-lingual reasoning compared to SFT.
- **Key Paradigm Shift:** Non-English RL training can be more effective for RL than English-centric RL approaches.
- **Future Directions:** Guidance for more equitable and effective multilingual reasoning development.

Thanks for listening !

Beyond English-Centric Training: How Reinforcement Learning Improves Cross-Lingual Reasoning

Shulin Huang, Yiran Ding, Junshu Pan, Yue Zhang*

huangshulin@westlake.edu.cn

Zhejiang University, Westlake University