

Multilingual Reinforcement Learning

Wei Xu
Associate professor
College of Computing, Machine Learning Center
Georgia Institute of Technology

𝕏 @cocoweixu | https://cocoxu.github.io/

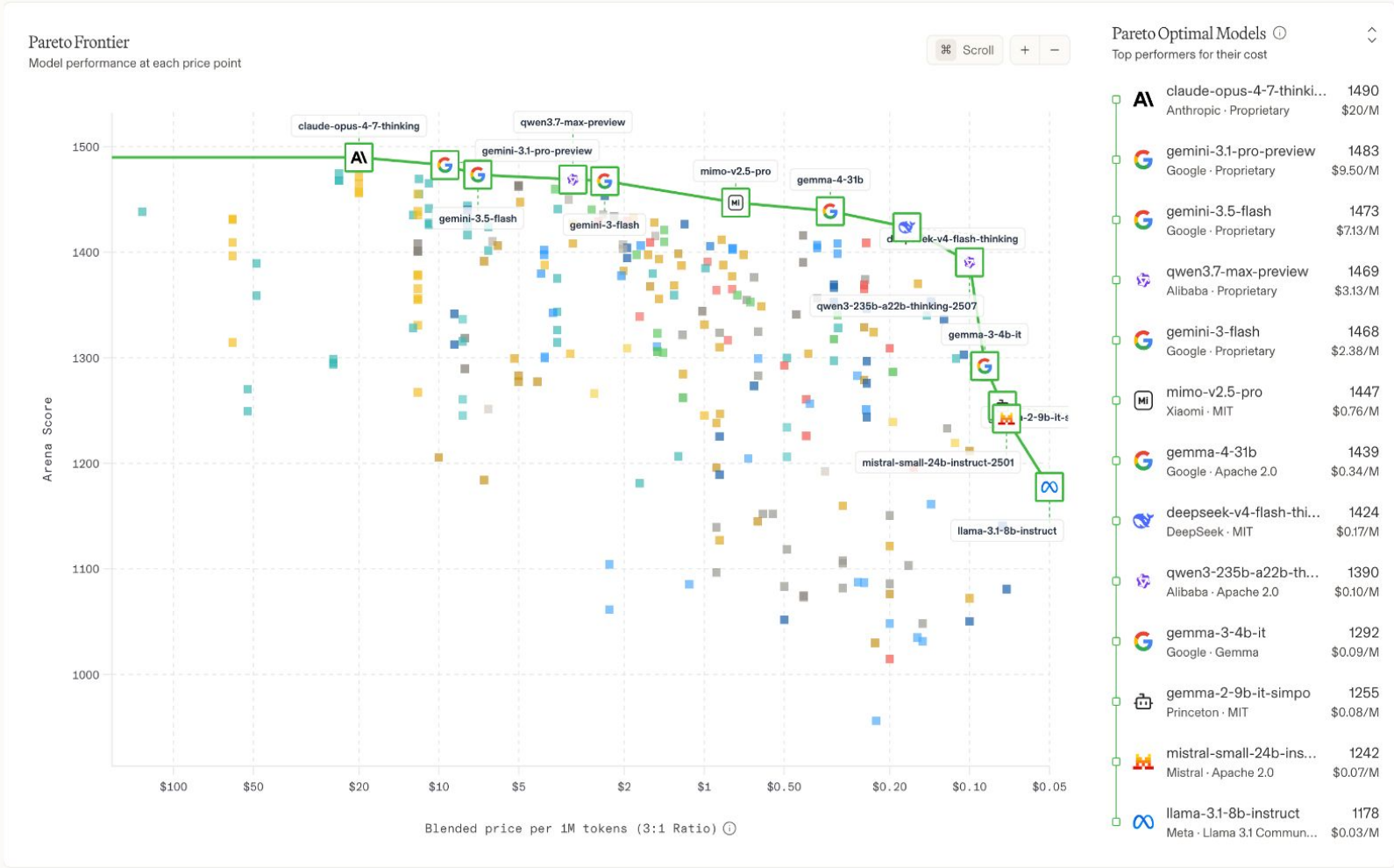
Which LLMs work better for non-English languages?

Text Arena | Non-English

View overall rankings across various AI models in text-to-text tasks across math, coding, creative writing, and other open-ended domains.

Jun 4, 2026 | 3,199,084 votes | 365 models

https://arena.ai/



LLMs perform better in English than in other languages*

Accuracy (↑) on mGSM-Rev2 dataset [Peter et al., 2025]

System	en	bn	de	es	fr	ja	ru	sw	te	th	zh
Gemini 2.5 Pro	99.6	98.8	100.0	99.6	98.8	100.0	99.6	99.2	99.2	100.0	99.6
Gemini 2.5 Flash	99.6	97.2	99.6	99.6	98.8	98.4	99.6	98.0	98.4	97.6	98.8
Gemma3 27B	100.0	91.2	97.2	97.2	98.0	93.2	96.8	91.6	93.6	95.2	95.2
Gemma3 12B	97.2	85.6	93.6	94.8	93.2	87.6	94.4	85.2	90.4	89.6	89.6
GPT5	99.6	99.2	99.6	98.8	98.4	100.0	99.6	98.4	99.6	98.8	100.0
GPT5 Nano	97.6	93.2	97.2	98.8	96.0	95.6	96.8	93.2	90.0	92.0	98.4
GPT OSS 20B	99.6	94.8	98.8	98.4	97.6	95.2	98.8	81.2	98.0	95.6	99.2
Claude Sonnet 4.5	99.6	99.2	99.6	99.2	100.0	99.2	100.0	98.4	98.4	98.0	99.2
Claude Haiku 4.5	99.6	98.8	98.8	98.4	98.8	99.2	99.2	96.8	98.8	97.6	98.4
DeepSeek R1	99.2	95.6	98.0	98.0	98.4	98.0	98.8	96.4	94.4	98.0	98.4
DeepSeek V3	100.0	96.4	99.6	99.2	98.8	97.6	98.0	95.2	94.4	98.0	99.6

a revised version of Multilingual Grade School Math (mGSM) [Shi et al., 2022] benchmark w/ 250 math problems translated from English GSM8K [Cobbe et al., 2021] into 10 languages

Jan-Thorsten Peter, David Vilar, Tobias Domhan, Dan Malkin, Markus Freitag.
"Mind the Gap... or Not? How Translation Errors and Evaluation Details Skew Multilingual Results" (arXiv 2025)

MGSM-Rev2: Multilingual Grade-school Math

125 tokens / 80 words / 339 chars

Artie has a flower stand at the Farmers Market. He sells three kinds of flowers: marigolds, petunias and begonias. He usually sells marigolds for $2.74 per pot, petunias for $1.87 per pot and begonias for $2.12 per pot. Artie has no change today, so he has decided to round all his prices to the nearest dollar. If Artie sells 12 pots of marigolds, 9 pots of petunias and 17 pots of begonias, how much will he make？

MGSM-Rev2 [Peter et al., 2025]

Language	Code	Alphabet	Resources
English	en	Latin	High
Bengali	bn	Bangla	Low
German	de	Latin	High
Spanish	es	Latin	High
French	fr	Latin	High
Japanese	ja	Kanji + Kana	Medium
Russian	ru	Cyrillic	Medium
Swahili	sw	Latin	Low
Telugu	te	Brahmic	Low
Thai	th	Thai	Low
Chinese	zh	Chinese	High

a revised MGSM [Shi et al., 2022] benchmark w/ 250 grade-school math problems translated from English GSM8K [Cobbe et al., 2021] into 10 languages

Jan-Thorsten Peter, David Vilar, Tobias Domhan, Dan Malkin, Markus Freitag. "Mind the Gap... or Not? How Translation Errors and Evaluation Details Skew Multilingual Results" (arXiv 2025)

MGSM-Rev2: Multilingual Grade-school Math

125 tokens / 80 words / 339 chars

Artie has a flower stand at the Farmers Market. He sells three kinds of flowers: marigolds, petunias and begonias. He usually sells marigolds for $2.74 per pot, petunias for $1.87 per pot and begonias for $2.12 per pot. Artie has no change today, so he has decided to round all his prices to the nearest dollar. If Artie sells 12 pots of marigolds, 9 pots of petunias and 17 pots of begonias, how much will he make?

288 tokens / 69 words / 464 chars

ఆర్టీకి రైతు మార్కెట్ వద్ద పువ్వుల స్టాండ్ ఉంది. అతడు మూడు రకాలైన పువ్వులు అమ్ముతాడు: బంతిపూలు, పెటూనియాస్ మరియు బెగోనియాలు. అతడు ...

128 tokens / 58 words / 137 chars

亚提在农贸市场有一个花卉摊位。他卖三种花：万寿菊、矮牵牛花和秋海棠。...

MGSM-Rev2 [Peter et al., 2025]

Language	Code	Alphabet	Resources
English	en	Latin	High
Bengali	bn	Bangla	Low
German	de	Latin	High
Spanish	es	Latin	High
French	fr	Latin	High
Japanese	ja	Kanji + Kana	Medium
Russian	ru	Cyrillic	Medium
Swahili	sw	Latin	Low
Telugu	te	Brahmic	Low
Thai	th	Thai	Low
Chinese	zh	Chinese	High

a revised MGSM [Shi et al., 2022] benchmark w/ 250 grade-school math problems translated from English GSM8K [Cobbe et al., 2021] into 10 languages

Jan-Thorsten Peter, David Vilar, Tobias Domhan, Dan Malkin, Markus Freitag. "Mind the Gap... or Not? How Translation Errors and Evaluation Details Skew Multilingual Results" (arXiv 2025)

LLMs perform better in English than in other languages*

By training models using RL with verifiable rewards, LLMs learn to do surprisingly effective reasoning, e.g., for Math.

But, this often works best in English.

DeepSeekMath - GRPO [Shao et al., 2024]

arXiv:2402.03300v3 [cs.CL] 27 Apr 2024

deepseek

DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models

Zhihong Shao[1,2†], Peiyi Wang[1,3†], Qihao Zhu[1,3†], Runxin Xu[1], Junxiao Song[1]
Xiao Bi[1], Haowei Zhang[1], Mingchuan Zhang[1], Y.K. Li[1], Y. Wu[1], Daya Guo[1*]

[1]DeepSeek-AI, [2]Tsinghua University, [3]Peking University

{zhihongshao,wangpeiyi,zhuqh,guoday}@deepseek.com
https://github.com/deepseek-ai/DeepSeek-Math

Abstract

Mathematical reasoning poses a significant challenge for language models due to its complex and structured nature. In this paper, we introduce DeepSeekMath 7B, which continues pre-training DeepSeek-Coder-Base-v1.5 7B with 120B math-related tokens sourced from Common Crawl, together with natural language and code data. DeepSeekMath 7B has achieved an impressive score of 51.7% on the competition-level MATH benchmark without relying on external toolkits and voting techniques, approaching the performance level of Gemini-Ultra and GPT-4. Self-consistency over 64 samples from DeepSeekMath 7B achieves 60.9% on MATH. The mathematical reasoning capability of DeepSeekMath is attributed to two key factors: First, we harness the significant potential of publicly available web data through a meticulously engineered data selection pipeline. Second, we introduce Group Relative Policy Optimization (GRPO), a variant of Proximal Policy Optimization (PPO), that enhances mathematical reasoning abilities while concurrently optimizing the memory usage of PPO.

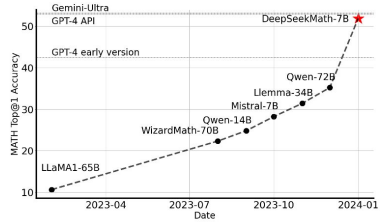


Figure 1 | Top1 accuracy of open-source models on the competition-level MATH benchmark (Hendrycks et al., 2021) without the use of external toolkits and voting techniques.

* Core contributors.
† Work done during internship at DeepSeek-AI.

But, how well do LLMs understand different cultures?

Common Sense

在中国，电影票上的放映时间是指电影正式开始的时间，还是说会先放预告片？
[In China, does the showtime on the ticket mean the movie starts then, or are there trailers first?]

Literacy

ما معنى "وجه ما بيضحك للرغيف السخن"؟
[What does “his face doesn’t smile to hot bread” mean?]

Social Norms

送40块钱的红包好不好？
[Is it okay to put ¥40 in a red envelope?]

🧧 Red envelope is a typical gift in China that contains cash

Opinions

アメリカの病院にかかっておどろくことは？
[What surprises you when visiting a hospital in the U.S.?]

Cultural Entities

墨西哥的国花是什么？
[What is Mexico’s national flower?]

3490 culture-specific Questions
x 10 Answers with human ratings
Chinese/Japanese/Arabic

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, Wei Xu
“🧡🌍CARE: Multilingual Human Preference Learning for Cultural Awareness” (EMNLP 2025)

CARE: Cultural Alignment from Human Preferences

A: The pronunciations of words “4” and "death" are similar in Chinese, so ¥40 is inappropriate as gift.

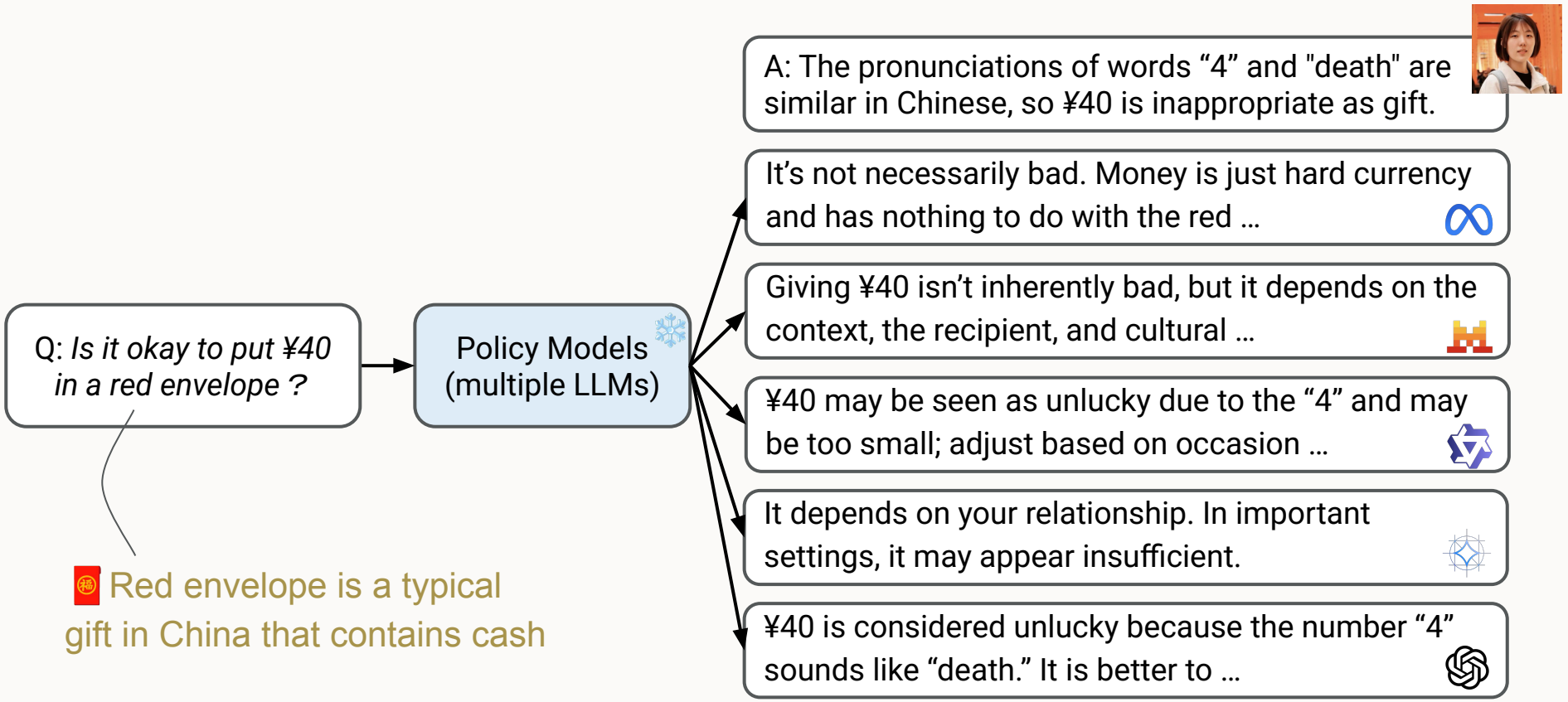


Q: *Is it okay to put ¥40 in a red envelope ?*

🧧 Red envelope is a typical gift in China that contains cash

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, Wei Xu
“CARE: Multilingual Human Preference Learning for Cultural Awareness” (EMNLP 2025)

CARE: Cultural Alignment from Human Preferences



Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, Wei Xu
“CARE: Multilingual Human Preference Learning for Cultural Awareness” (EMNLP 2025)

CARE: Cultural Alignment from Human Preferences

Q: *Is it okay to put ¥40 in a red envelope ?*

Red envelope is a typical gift in China that contains cash

Policy Models (multiple LLMs)

A: The pronunciations of words “4” and "death" are similar in Chinese, so ¥40 is inappropriate as gift.



It’s not necessarily bad. Money is just hard currency and has nothing to do with the red …

Giving ¥40 isn’t inherently bad, but it depends on the context, the recipient, and cultural …

¥40 may be seen as unlucky due to the “4” and may be too small; adjust based on occasion …

It depends on your relationship. In important settings, it may appear insufficient.

¥40 is considered unlucky because the number “4” sounds like “death.” It is better to …

rated by humans on 1-10 scale (↑)

10

1

2

7

4

9

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, Wei Xu
“CARE: Multilingual Human Preference Learning for Cultural Awareness” (EMNLP 2025)

CARE: Cultural Alignment from Human Preferences

meaning … these collected outputs will not change during training
(i.e., offline reinforcement learning)

These models are “frozen”

Q: *Is it okay to put ¥40 in a red envelope ?*

Policy Models (multiple LLMs)

A: The pronunciations of words “4” and "death" are similar in Chinese, so ¥40 is inappropriate as gift.

It’s not necessarily bad. Money is just hard currency and has nothing to do with the red …

Giving ¥40 isn’t inherently bad, but it depends on the context, the recipient, and cultural …

¥40 may be seen as unlucky due to the “4” and may be too small; adjust based on occasion …

It depends on your relationship. In important settings, it may appear insufficient.

¥40 is considered unlucky because the number “4” sounds like “death.” It is better to …

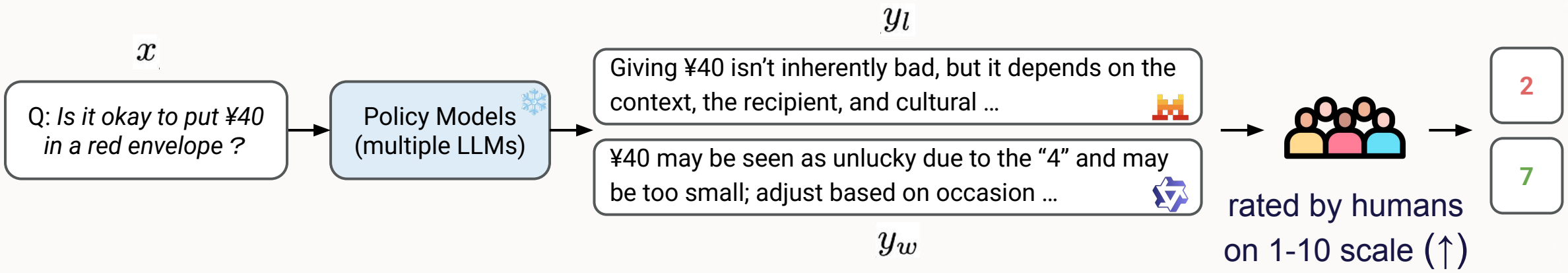
rated by humans on 1-10 scale (↑)

10
1
2
7
4
9

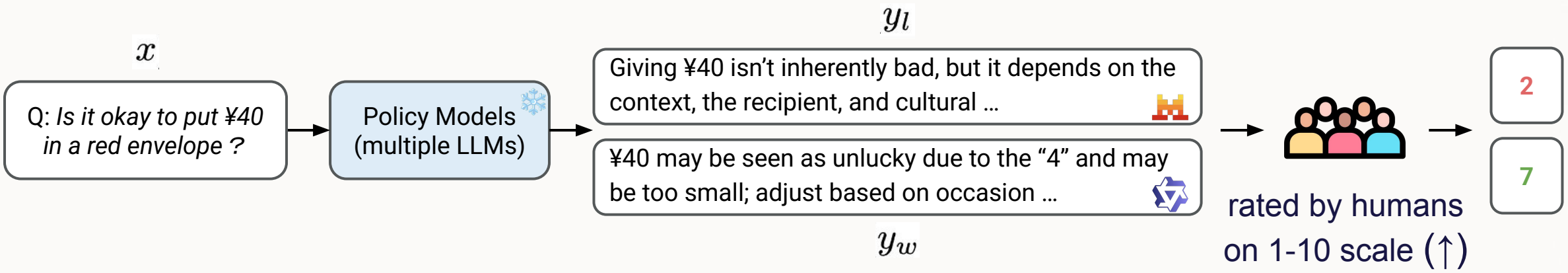
Red envelope is a typical gift in China that contains cash

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, Wei Xu
“CARE: Multilingual Human Preference Learning for Cultural Awareness” (EMNLP 2025)

CARE: Cultural Alignment from Human Preferences



CARE: Cultural Alignment from Human Preferences



Direct Preference Optimization [Rafailov et al., 2022]

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right]$$

CARE: Cultural Alignment from Human Preferences

Kahneman-Tversky Optimization [Ethayarajh et al., 2024]

$$\mathcal{L}_{\text{KTO}} = -\lambda_w \, \sigma\Big(\beta\big[\log \pi_\theta(y_w \mid x) - \log \pi_{\text{ref}}(y_w \mid x)\big] - z_{\text{ref}}\Big) + \lambda_\ell \, \sigma\Big(z_{\text{ref}} - \beta\big[\log \pi_\theta(y_\ell \mid x) - \log \pi_{\text{ref}}(y_\ell \mid x)\big]\Big).$$

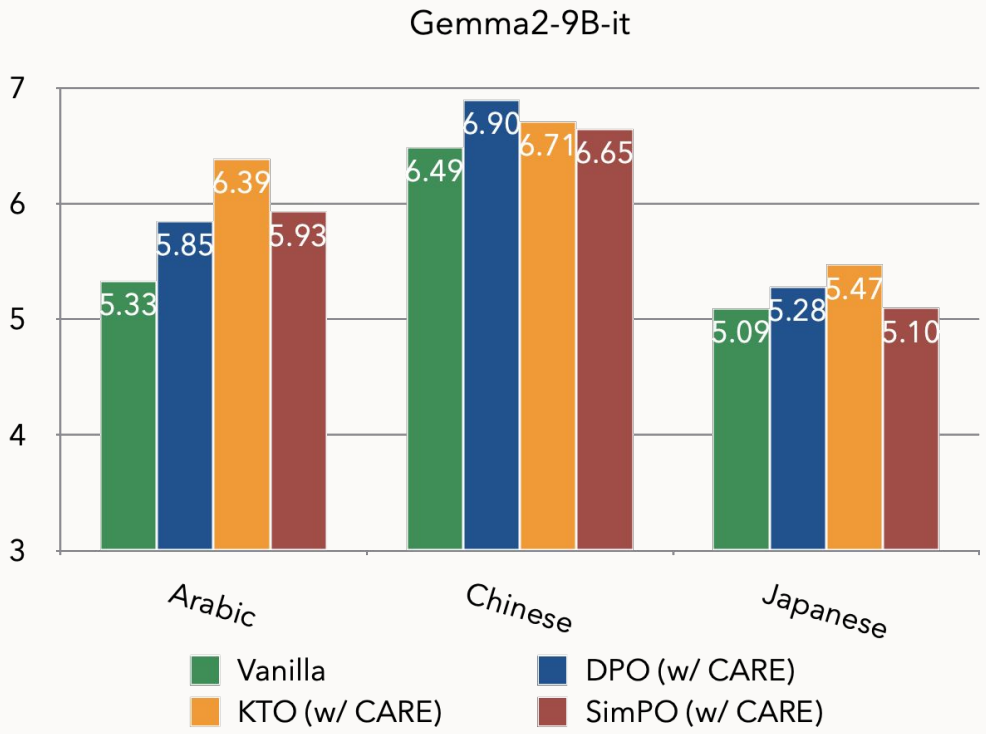
Simple Preference Optimization [Meng et al., 2024]

$$\mathcal{L}_{\text{SimPO}} = -\log \sigma\Big(\frac{\beta}{|y_w|} \log \pi_\theta(y_w \mid x) - \frac{\beta}{|y_\ell|} \log \pi_\theta(y_\ell \mid x) - \gamma\Big)$$

Direct Preference Optimization [Rafailov et al., 2022]

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]$$

CARE: Cultural Alignment from Human Preferences



Kahneman-Tversky Optimization [Ethayarajh et al., 2024]

$$\mathcal{L}_{\text{KTO}} = -\lambda_w \, \sigma\Big(\beta\big[\log \pi_\theta(y_w \mid x) - \log \pi_{\text{ref}}(y_w \mid x)\big] - z_{\text{ref}}\Big) + \lambda_\ell \, \sigma\Big(z_{\text{ref}} - \beta\big[\log \pi_\theta(y_\ell \mid x) - \log \pi_{\text{ref}}(y_\ell \mid x)\big]\Big).$$

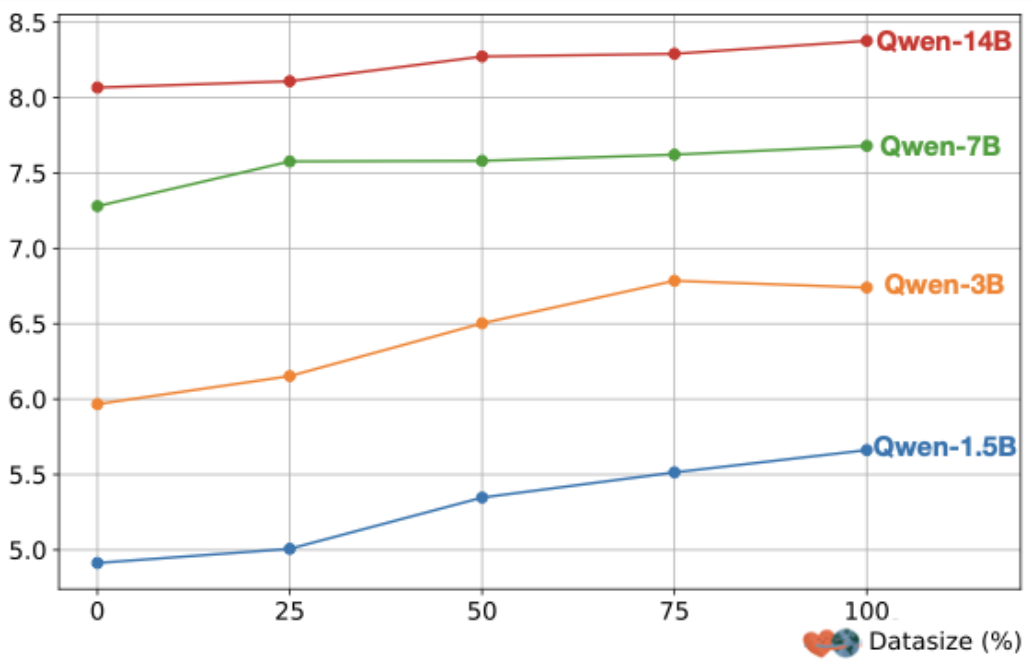
Simple Preference Optimization [Meng et al., 2024]

$$\mathcal{L}_{\text{SimPO}} = -\log \sigma\Big(\frac{\beta}{|y_w|} \log \pi_\theta(y_w \mid x) - \frac{\beta}{|y_\ell|} \log \pi_\theta(y_\ell \mid x) - \gamma\Big)$$

Direct Preference Optimization [Rafailov et al., 2022]

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l)\sim\mathcal{D}}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]$$

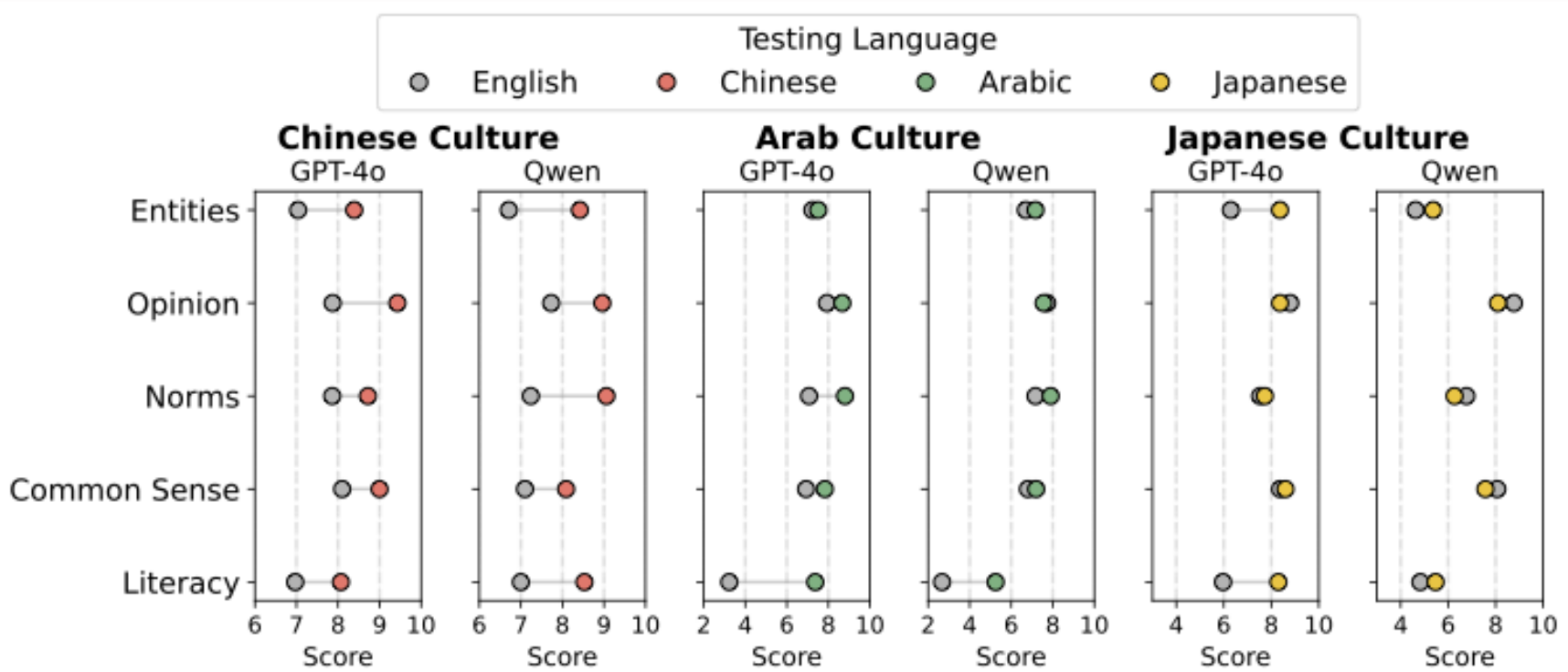
CARE: Cultural Alignment from Human Preferences



Consistent improvement
(DPO/SimPO/KTO)
across models w/ ~1000 Qs × 10 answers
rated by humans per langage

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, Wei Xu
"CARE: Multilingual Human Preference Learning for Cultural Awareness" (EMNLP 2025)

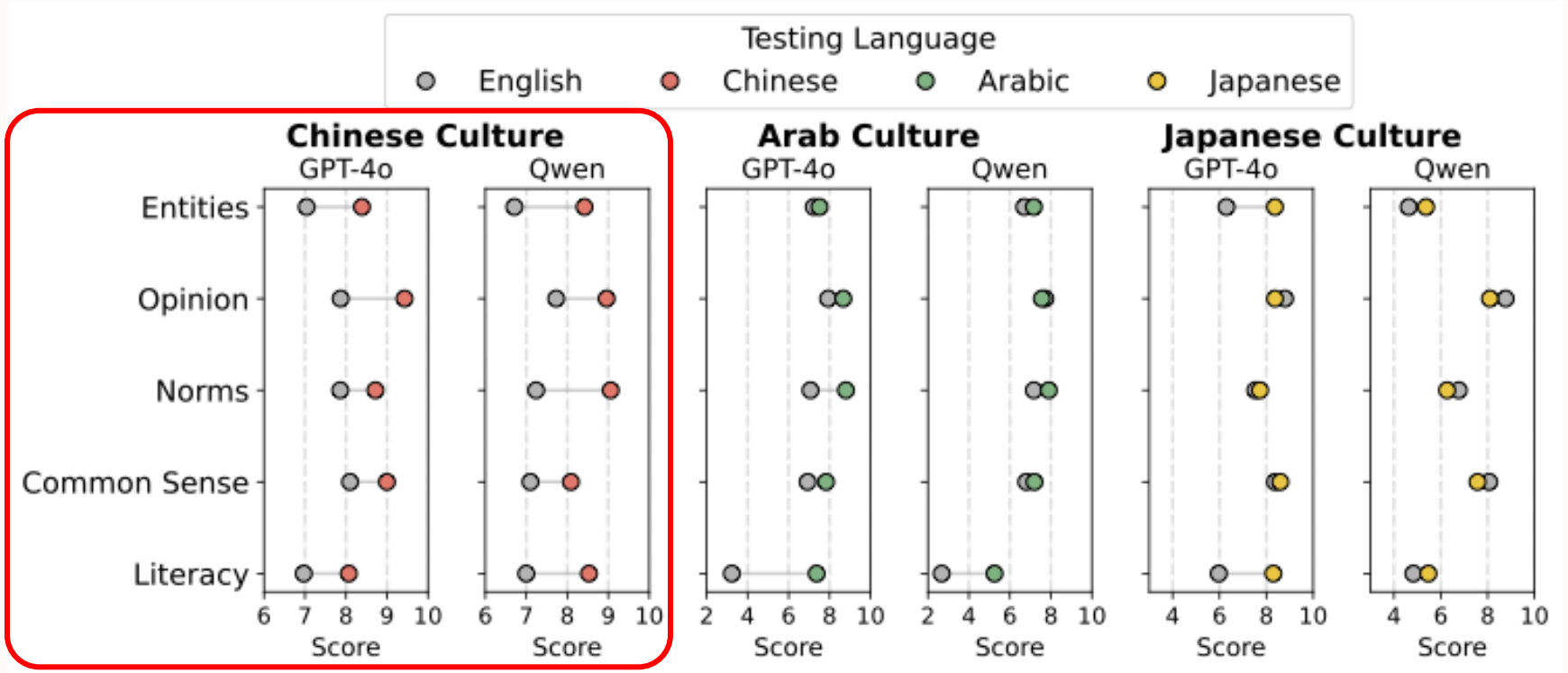
How well do LLMs understand different cultures?



Asking LLMs about Chinese culture in Chinese yields substantially better results than asking in English, while the language gap is much smaller for Arabic and Japanese cultures.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, Wei Xu
"CARE: Multilingual Human Preference Learning for Cultural Awareness" (EMNLP 2025)

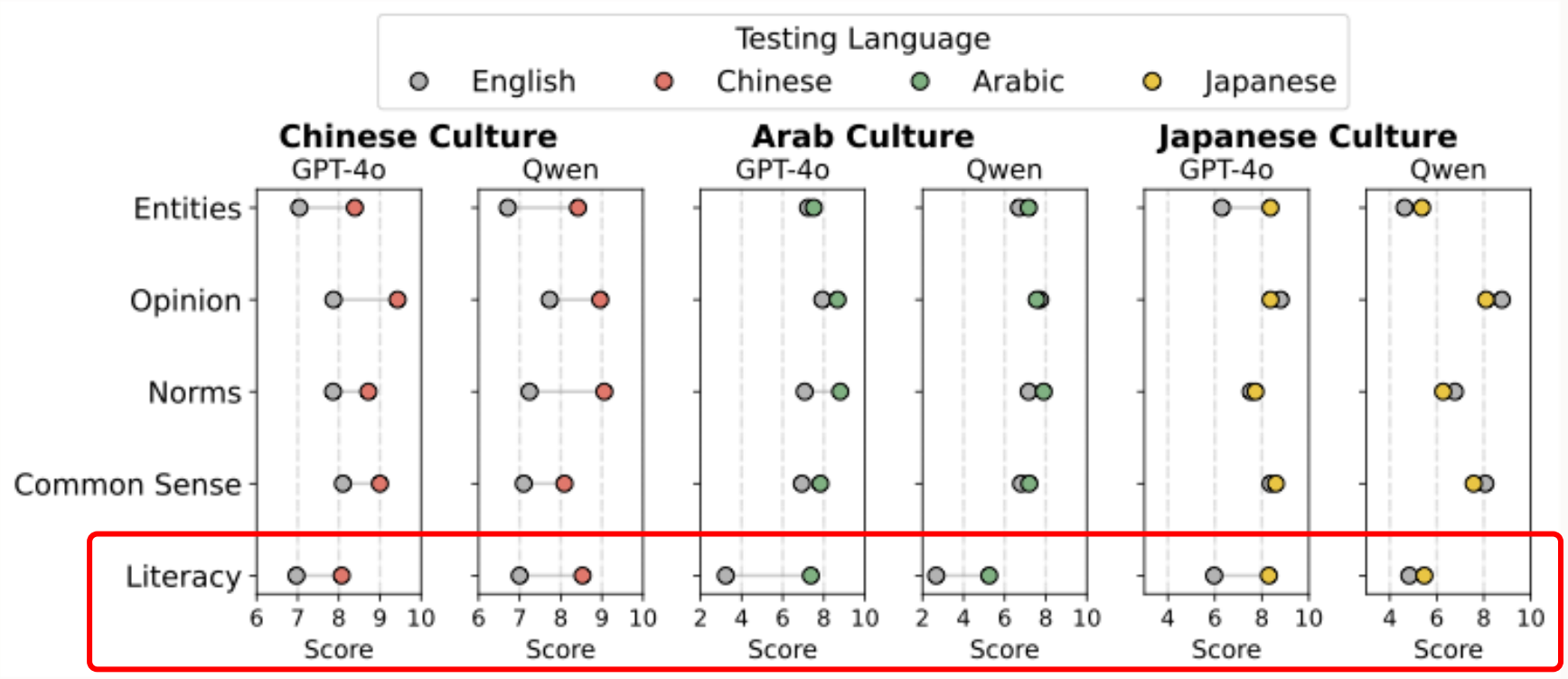
How well do LLMs understand different cultures?



Asking LLMs about Chinese culture in Chinese yields substantially better results than asking in English, while the language gap is much smaller for Arabic and Japanese cultures.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, Wei Xu
"CARE: Multilingual Human Preference Learning for Cultural Awareness" (EMNLP 2025)

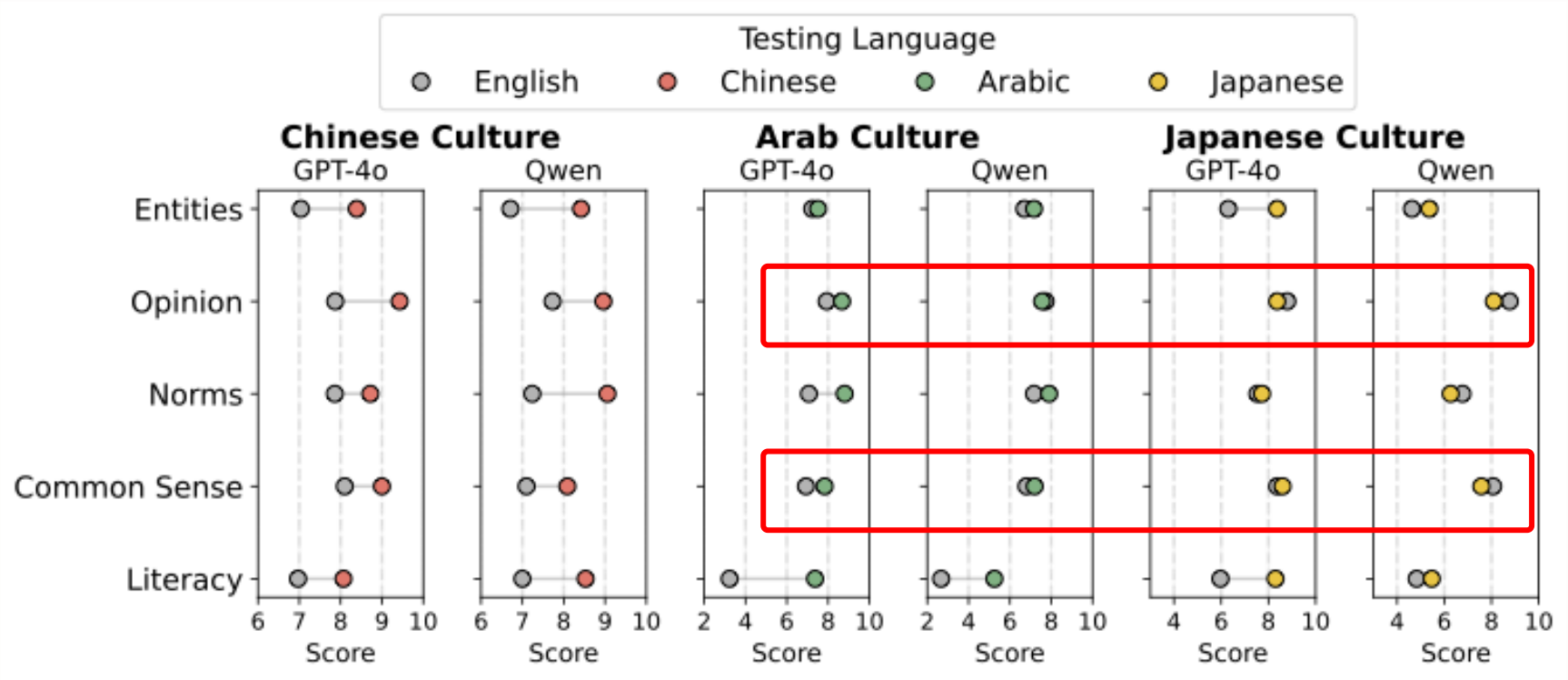
How well do LLMs understand different cultures?



Asking in native language also has the **biggest** advantage for **Literacy** related questions

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, Wei Xu
"CARE: Multilingual Human Preference Learning for Cultural Awareness" (EMNLP 2025)

How well do LLMs understand different cultures?



Asking in native language also has the **biggest** advantage for **Literacy** related questions, but the **least** advantage for cultural **Common Sense & Opinions**.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, Wei Xu
"CARE: Multilingual Human Preference Learning for Cultural Awareness" (EMNLP 2025)

How well do LLMs understand different cultures?





This may be partly because cultural differences, culture shock, and culturally specific common sense are often discussed from an outsider's perspective, making such information more prevalent in languages other than the culture's native language.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, Wei Xu
"CARE: Multilingual Human Preference Learning for Cultural Awareness" (EMNLP 2025)

So, that’s the backdrop

Learning from Multilingual Human Preference

offline reinforcement learning

Knowledge is not evenly distributed across languages.

Which leads us to ...

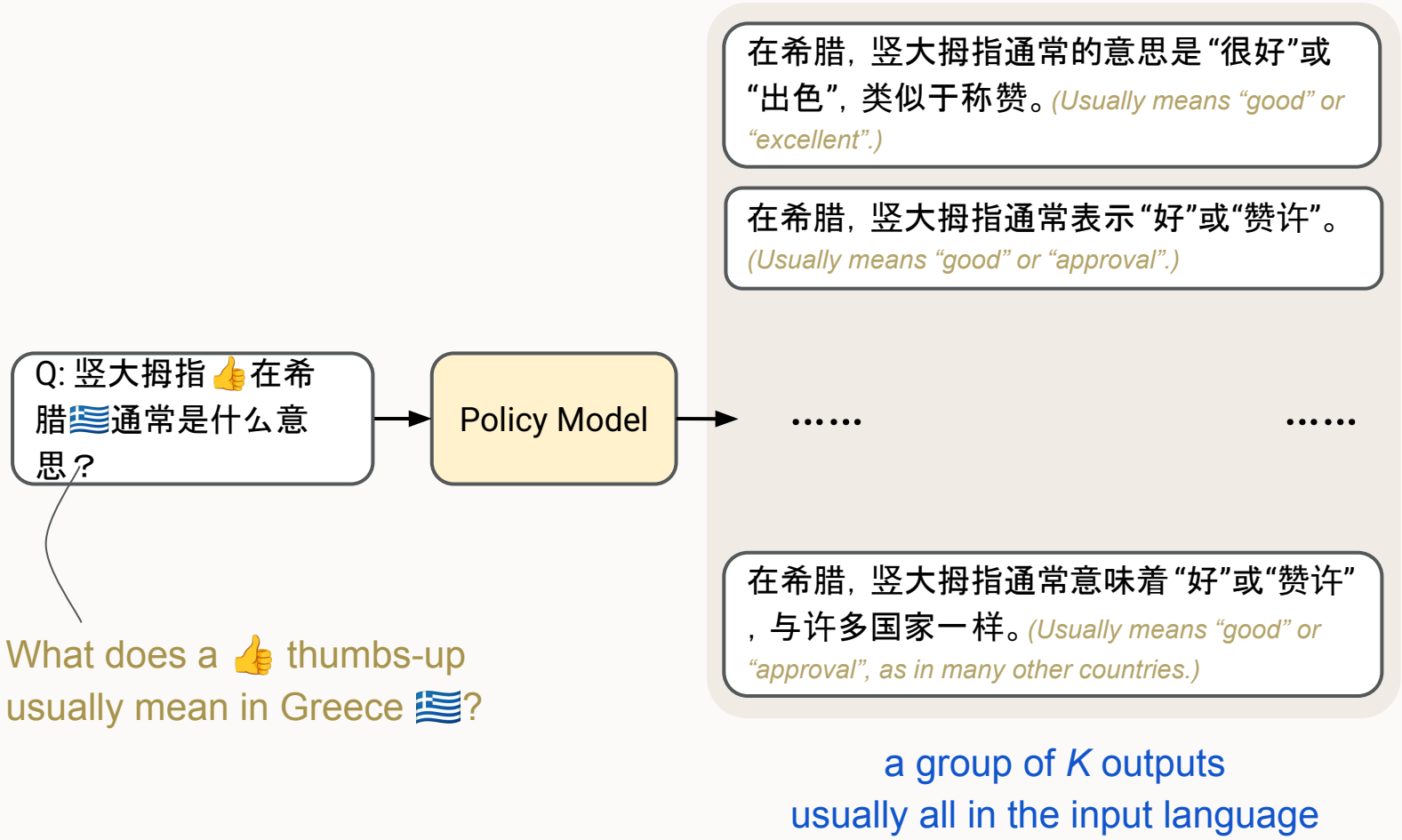
Learning from Multilingual Human Preference → **Multilingual Reinforcement Learning**

offline reinforcement learning → online reinforcement learning

Knowledge is not evenly distributed across languages.

GRPO: Group Relative Policy Optimization

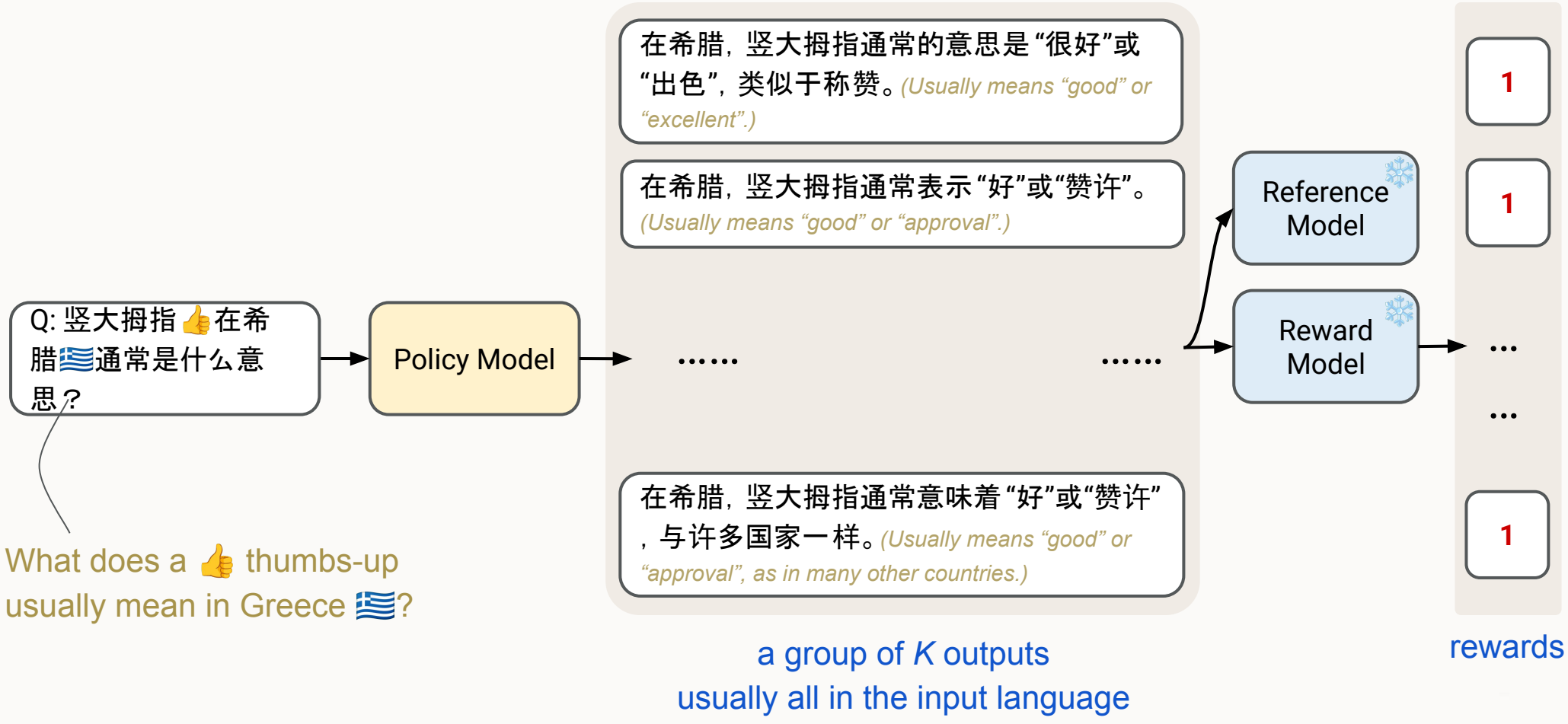
[Shao et al., 2025]



a group of *K* outputs
usually all in the input language

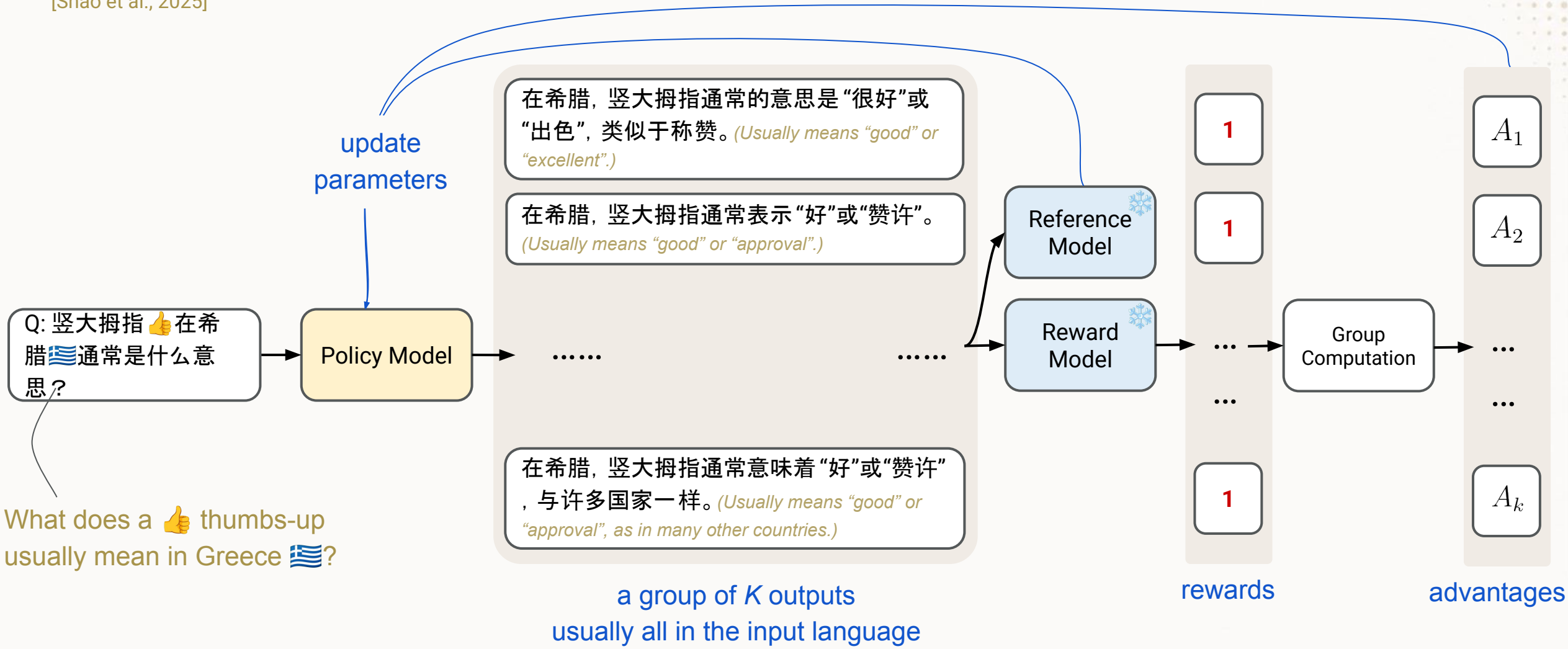
GRPO: Group Relative Policy Optimization

[Shao et al., 2025]



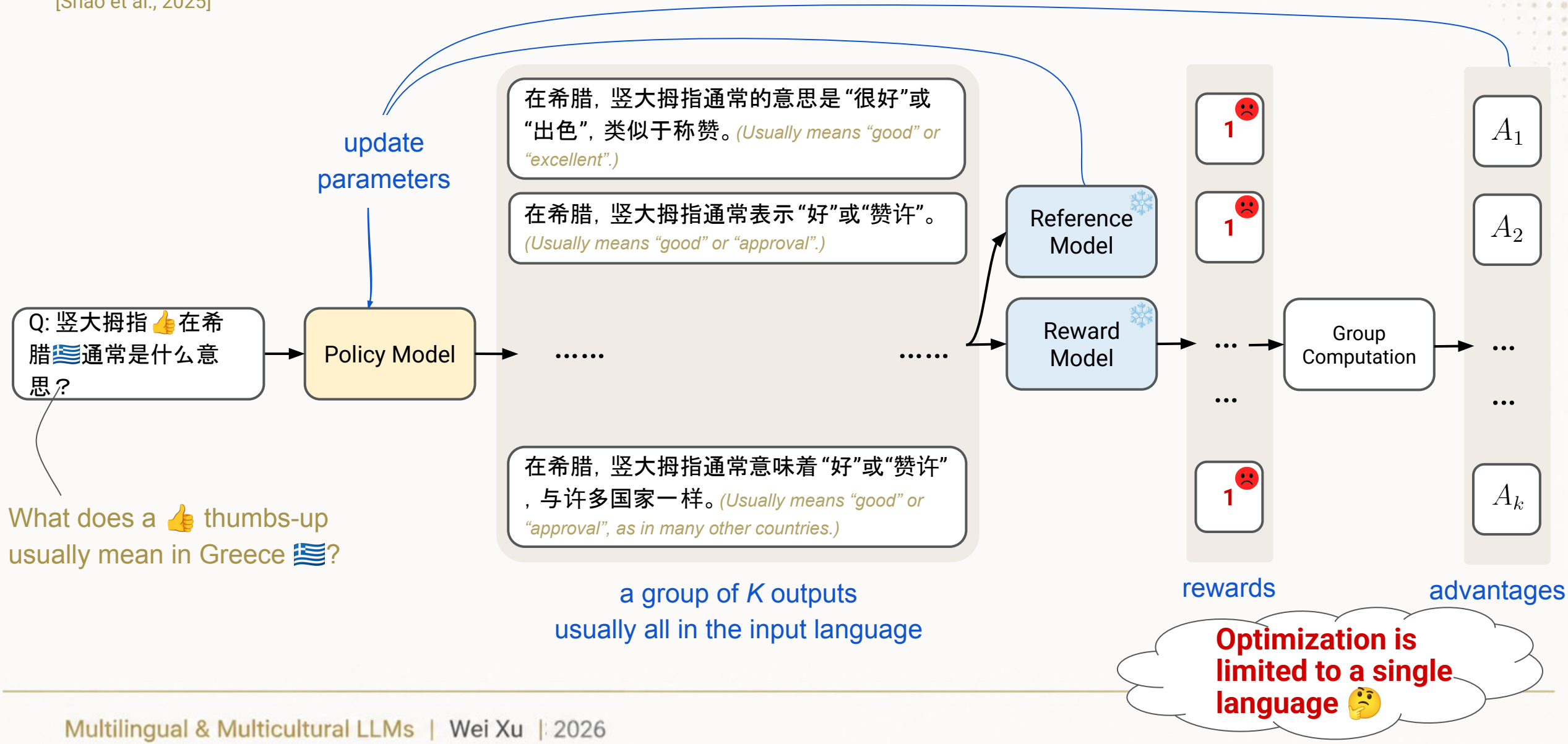
GRPO: Group Relative Policy Optimization

[Shao et al., 2025]

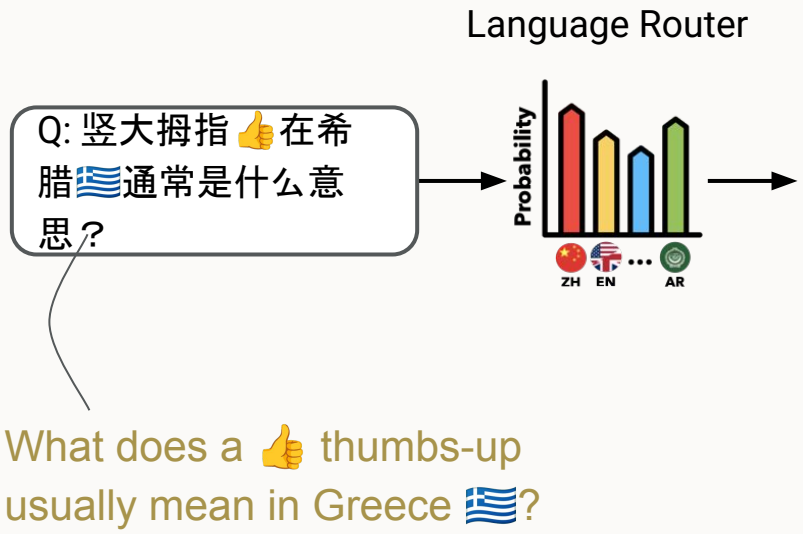


GRPO: Group Relative Policy Optimization

[Shao et al., 2025]

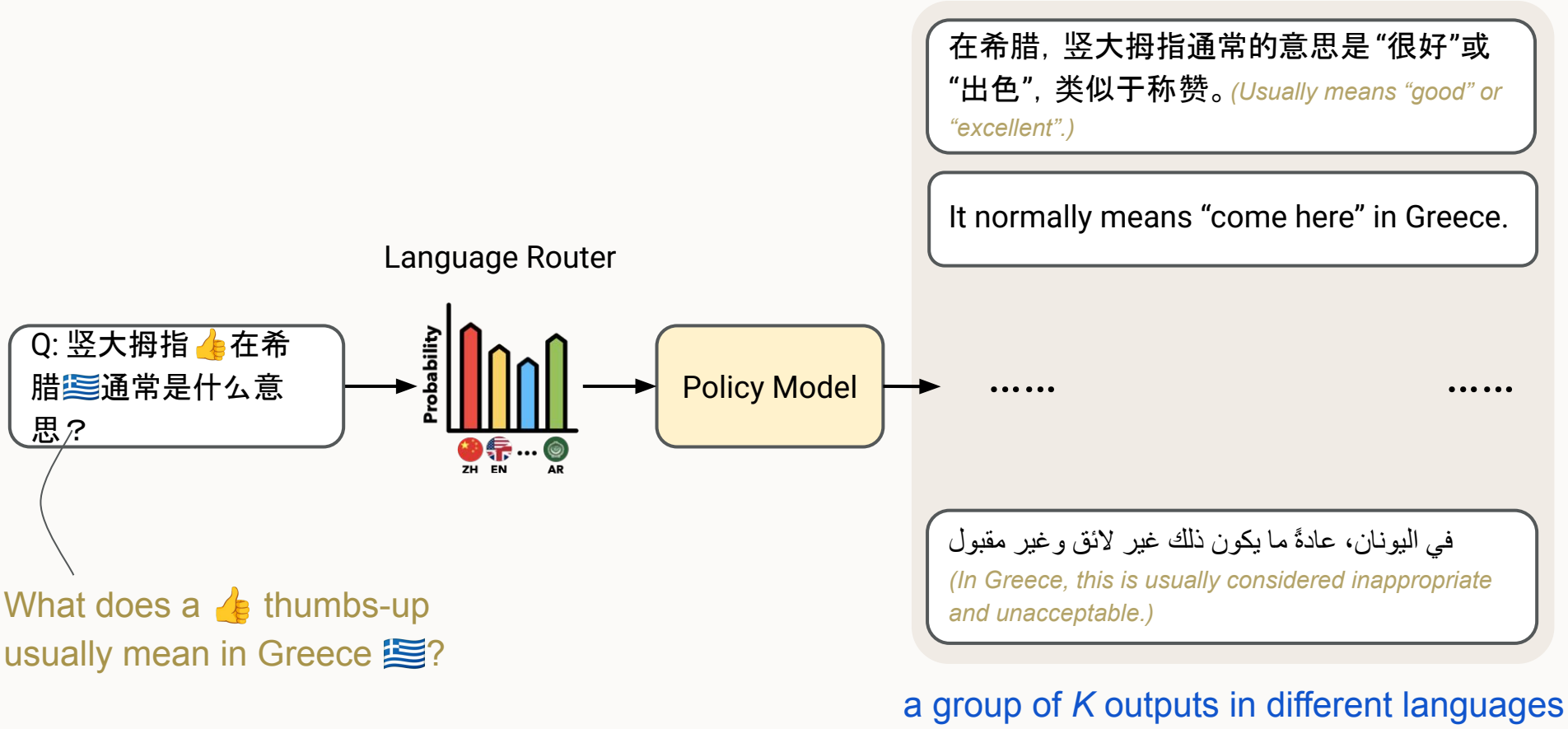


LRPO: Language-Routed Policy Optimization



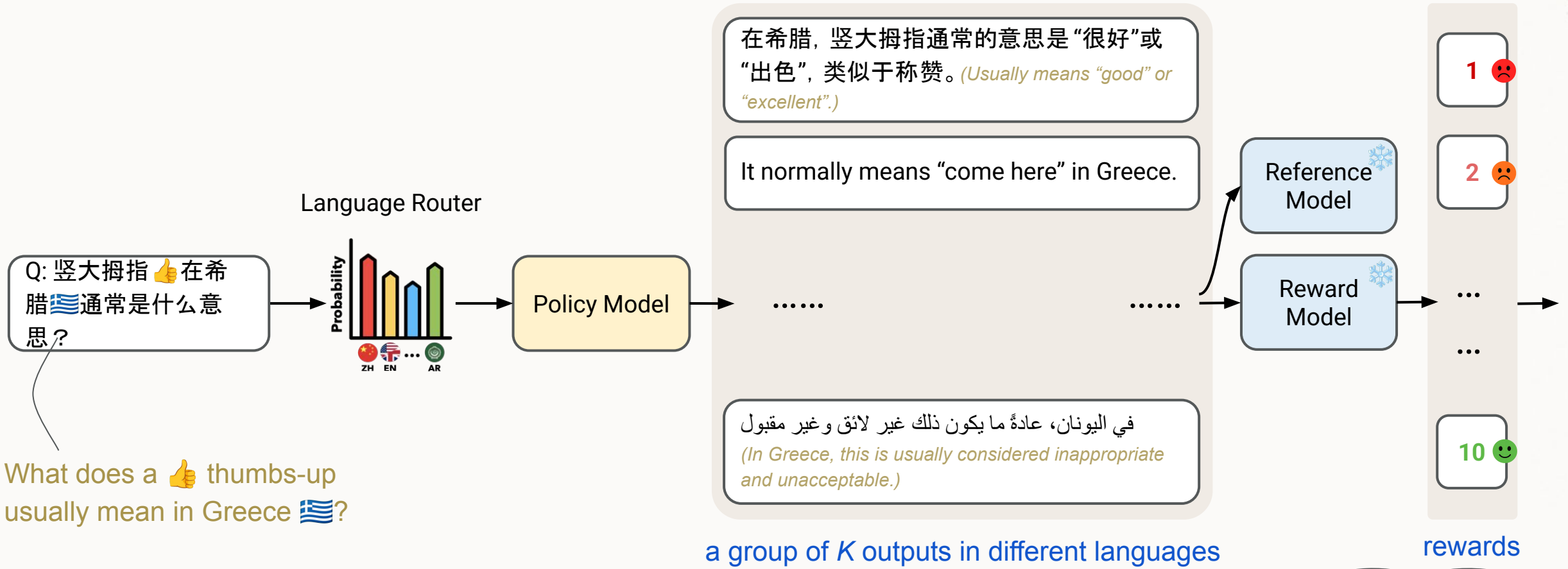
Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, Wei Xu
"Learning to Route Languages for Multilingual Policy Optimization" (ICML 2026)

LRPO: Language-Routed Policy Optimization

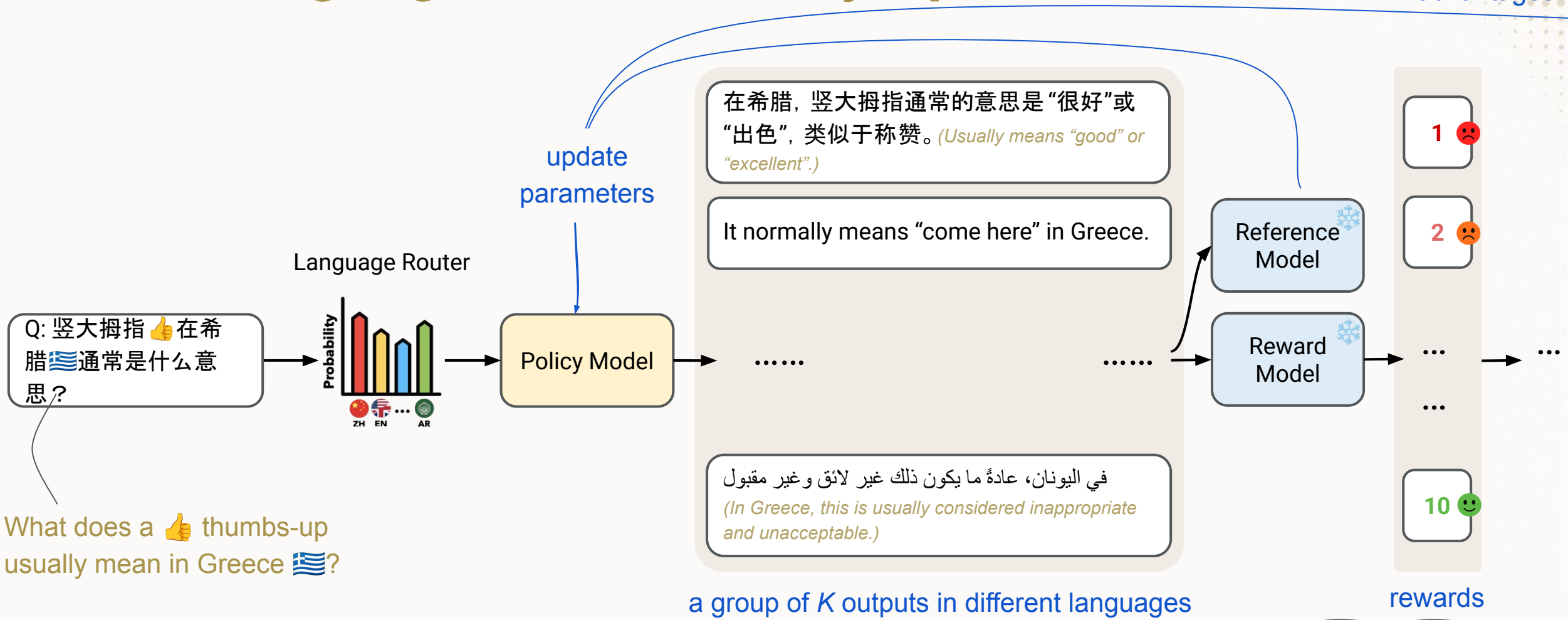


Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, Wei Xu
"Learning to Route Languages for Multilingual Policy Optimization" (ICML 2026)

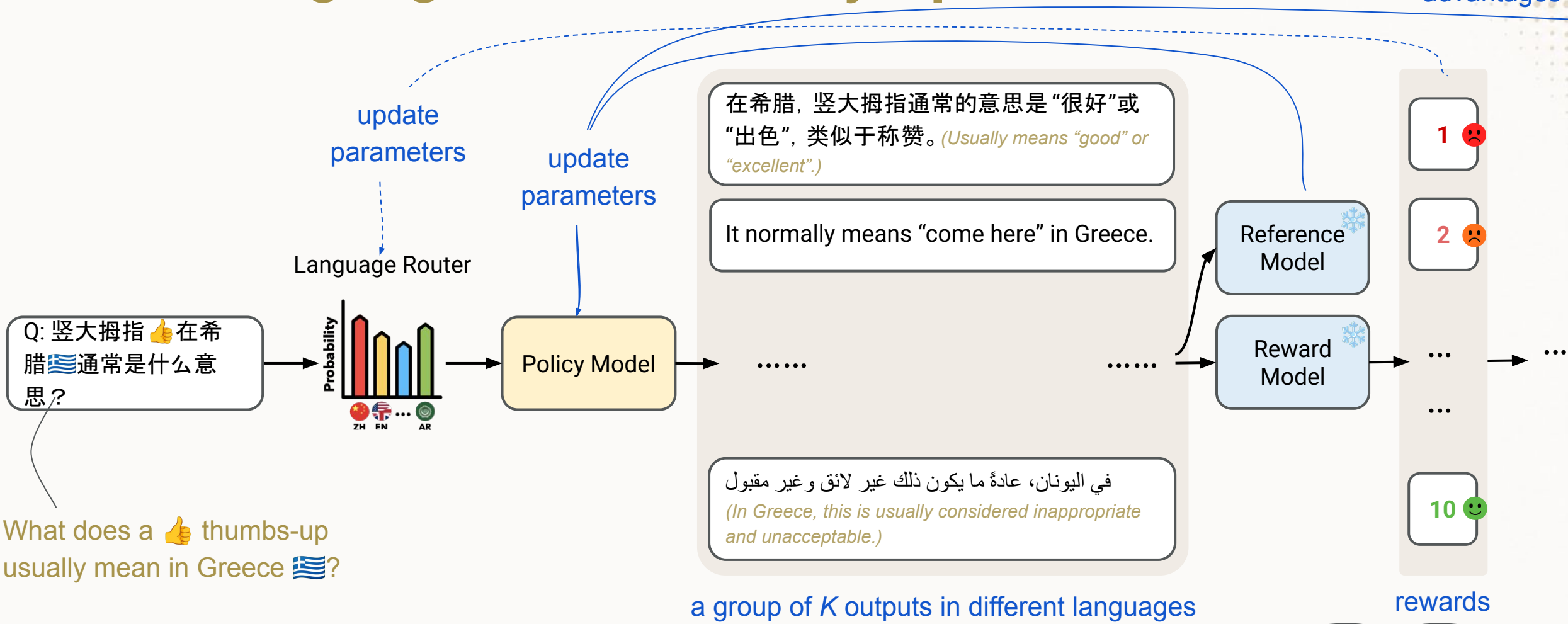
LRPO: Language-Routed Policy Optimization



LRPO: Language-Routed Policy Optimization

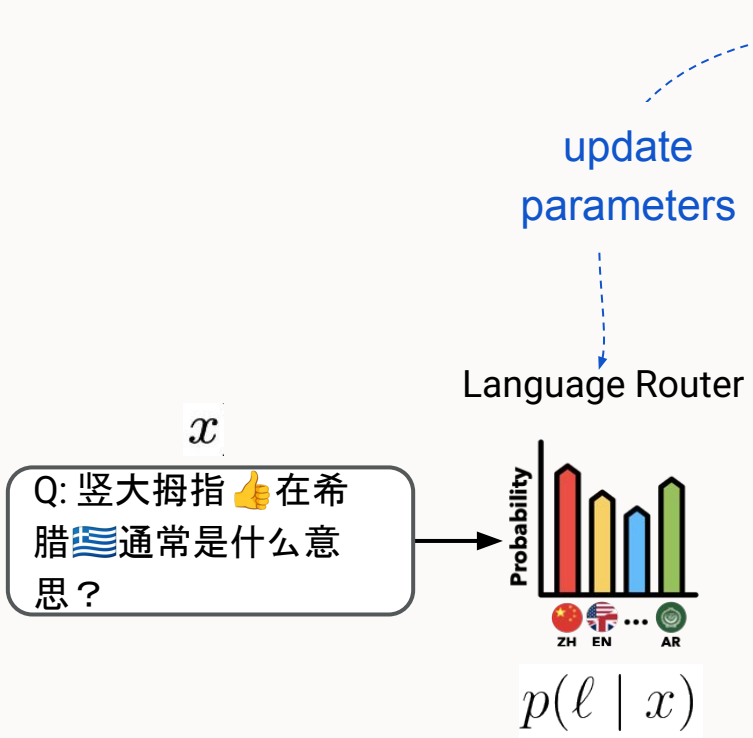


LRPO: Language-Routed Policy Optimization



The language router is trained jointly with the policy model using reward signals.

LRPO: Language-Routed Policy Optimization



💡 **A lightweight, adaptive language router that learns to select the more informative languages during RL training.**

Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, Wei Xu
"Learning to Route Languages for Multilingual Policy Optimization" (ICML 2026)

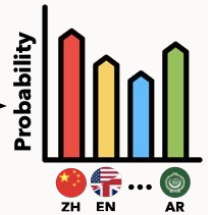
LRPO: Language-Routed Policy Optimization

update parameters

Language Router

x

Q: 竖大拇指👍在希腊🇬🇷通常是什么意思？



$$p(\ell \mid x)$$

$$\propto \exp\left(\frac{A_{t(x),\ell} + \mathbb{I}[g(x) \neq \varnothing]\, B_{g(x),\ell}}{\tau}\right)$$

topic-by-language matrix *A*

Topic	…	AR	EN	ZH	…
…					
General	…	0.3	0.9	0.8	…
Regional	…	0.4	0.5	0.6	…
…					

region-by-language matrix *B*

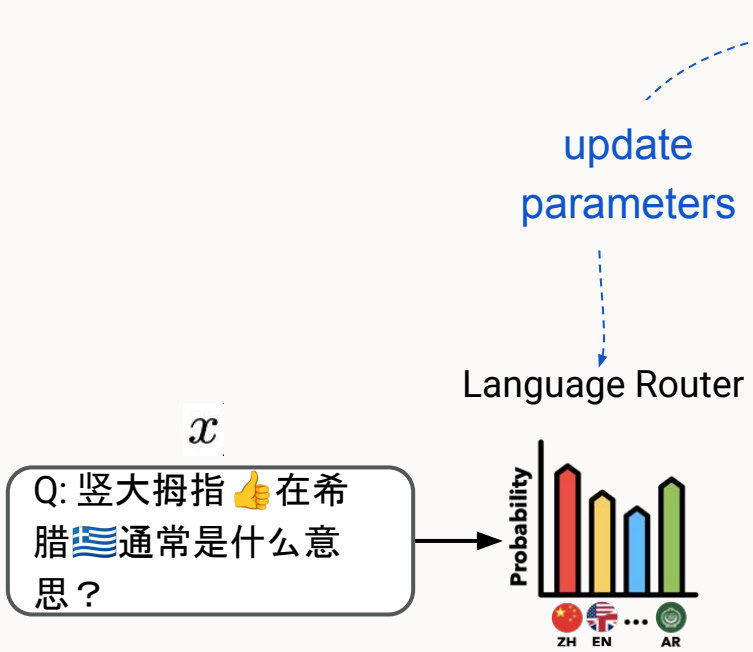
Region	…	AR	EN	ZH	…
…					
China	…	0.1	0.6	0.9	…
Greece	…	0.6	0.4	0.1	…
…					

The language router is parameterized by two matrices, *A* and *B*.

t - topic g- region ℓ- language

Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, Wei Xu
"Learning to Route Languages for Multilingual Policy Optimization" (ICML 2026)

LRPO: Language-Routed Policy Optimization



topic-by-language matrix *A*

Topic	…	AR	EN	ZH	…
…					
General	…	0.3	0.9	0.8	…
Regional	…	0.4	0.5	0.6	…
…					

region-by-language matrix *B*

Region	…	AR	EN	ZH	…
…					
China	…	0.1	0.6	0.9	…
Greece	…	0.6	0.4	0.1	…
…					

Training language router as a multi-armed bandit.

Action: choosing each language

Objective: maximize expected reward $\bar{r}$

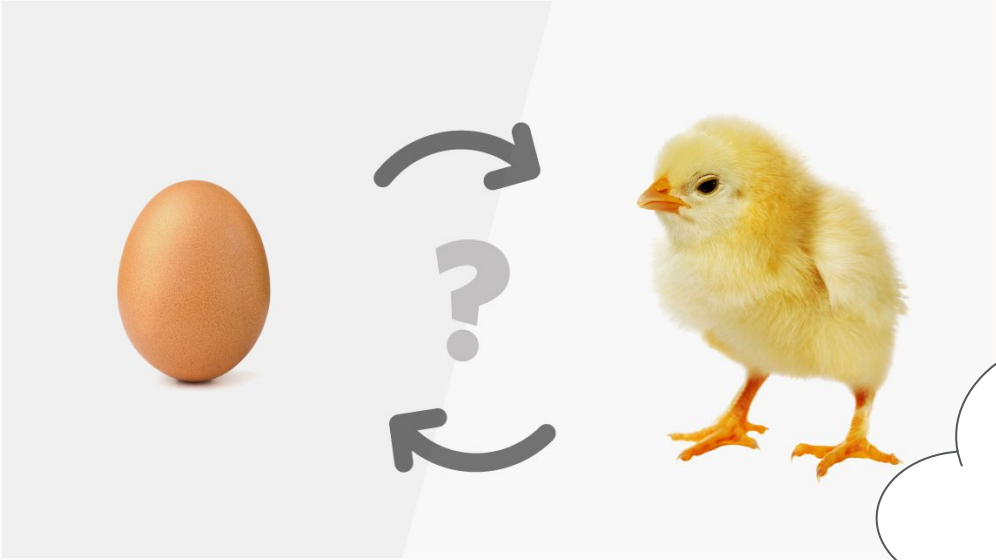
t - topic g- region ℓ- language

$$A_{t(x),\ell} \leftarrow (1-\alpha)\, A_{t(x),\ell} + \alpha\, \bar{r}_{t,g,\ell}$$
$$B_{g(x),\ell} \leftarrow (1-\alpha)\, B_{g(x),\ell} + \alpha\, \bar{r}_{t,g,\ell}$$

Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, Wei Xu
"Learning to Route Languages for Multilingual Policy Optimization" (ICML 2026)

Multilingual Reinforcement Learning

⚔️ **Main Challenge**: multilingual rewards are hard to get right



Languages needing the most improvement in LLM are also languages where rewards models are less reliable

Multilingual Reinforcement Learning

⚔️ **Main Challenge**: multilingual rewards are hard to get right

Approaches	Advantages	Limitations
RLVR (Reinforcement Learning with Verifiable Rewards) [polyGRPO - Wu et al., 2026]	Language-agnostic; stable	Only works for verifiable tasks (e.g., math, coding) ☹

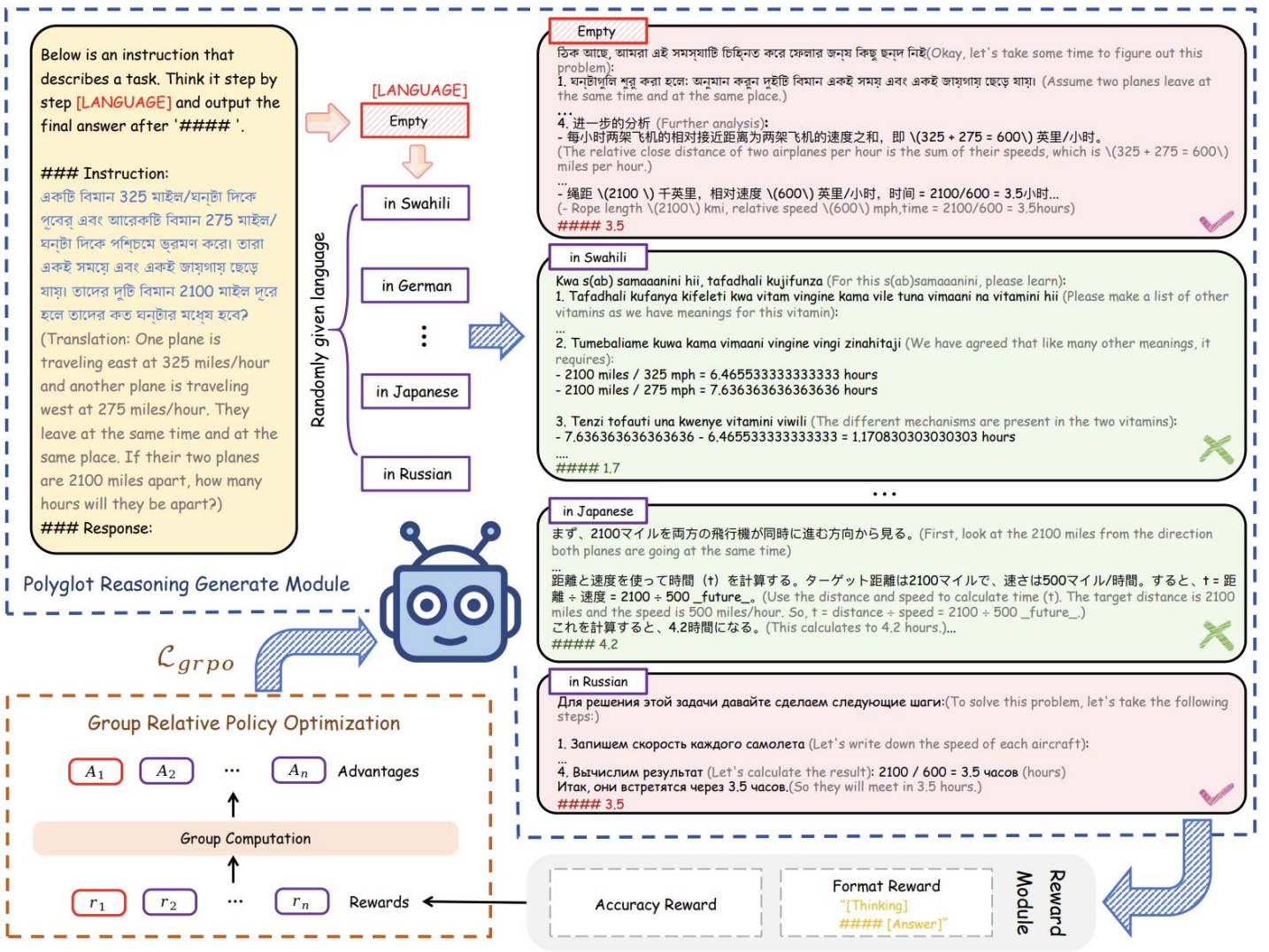
polyGRPO: Polyglot Group Relative Policy Optimization

Random Language for Rollouts

+

Verifiable Rewards

(format, accuracy for math Qs)



Linjuan Wu, Haoran Wei, Jialong Tang, Shuang Luo, Baosong Yang, Yongliang Shen, Weiming Lu "Language as a Latent Variable for Reasoning Optimization" (arXiv 2026)

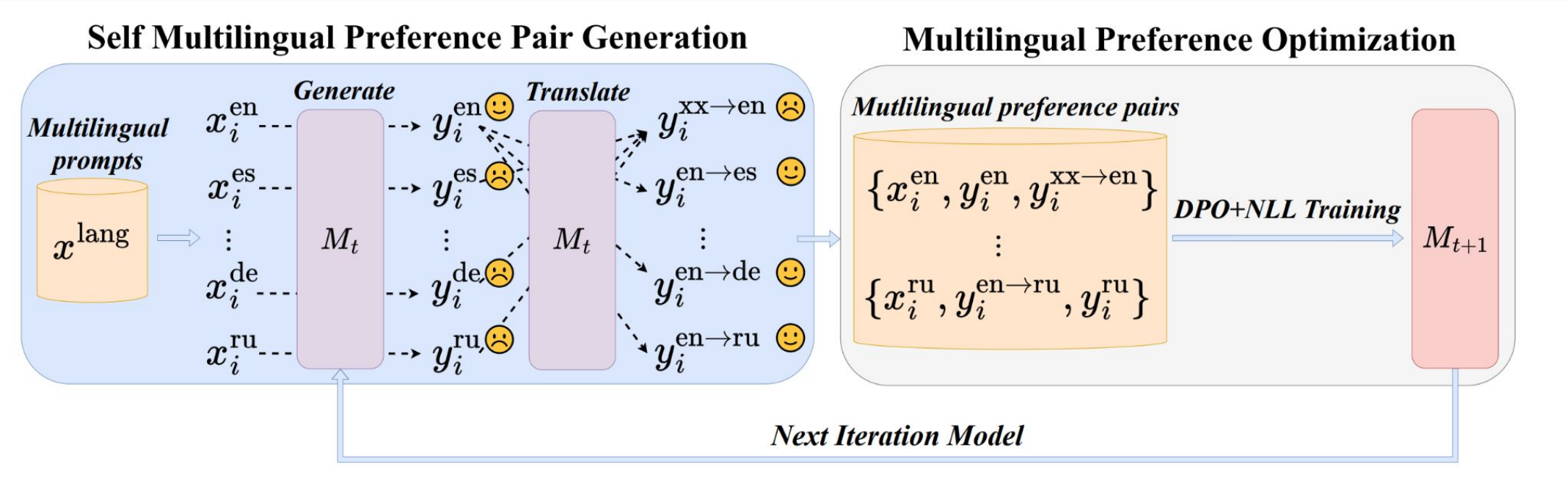
Multilingual Reinforcement Learning

⚔️ **Main Challenge**: multilingual rewards are hard to get right

Approaches	Advantages	Limitations
RLVR (Reinforcement Learning with Verifiable Rewards) [polyGRPO - Wu et al., 2026]	Language-agnostic; stable	Only works for verifiable tasks (e.g., math, coding) ☹
Pivot-Based Alignment [MAPO - She et al., 2024; MPO - Zhao et al., 2025]	Leverages strong English models; scalable	English-centric; may lose cultural/language nuances ☹

LIDR: Language Imbalance Driven Rewarding

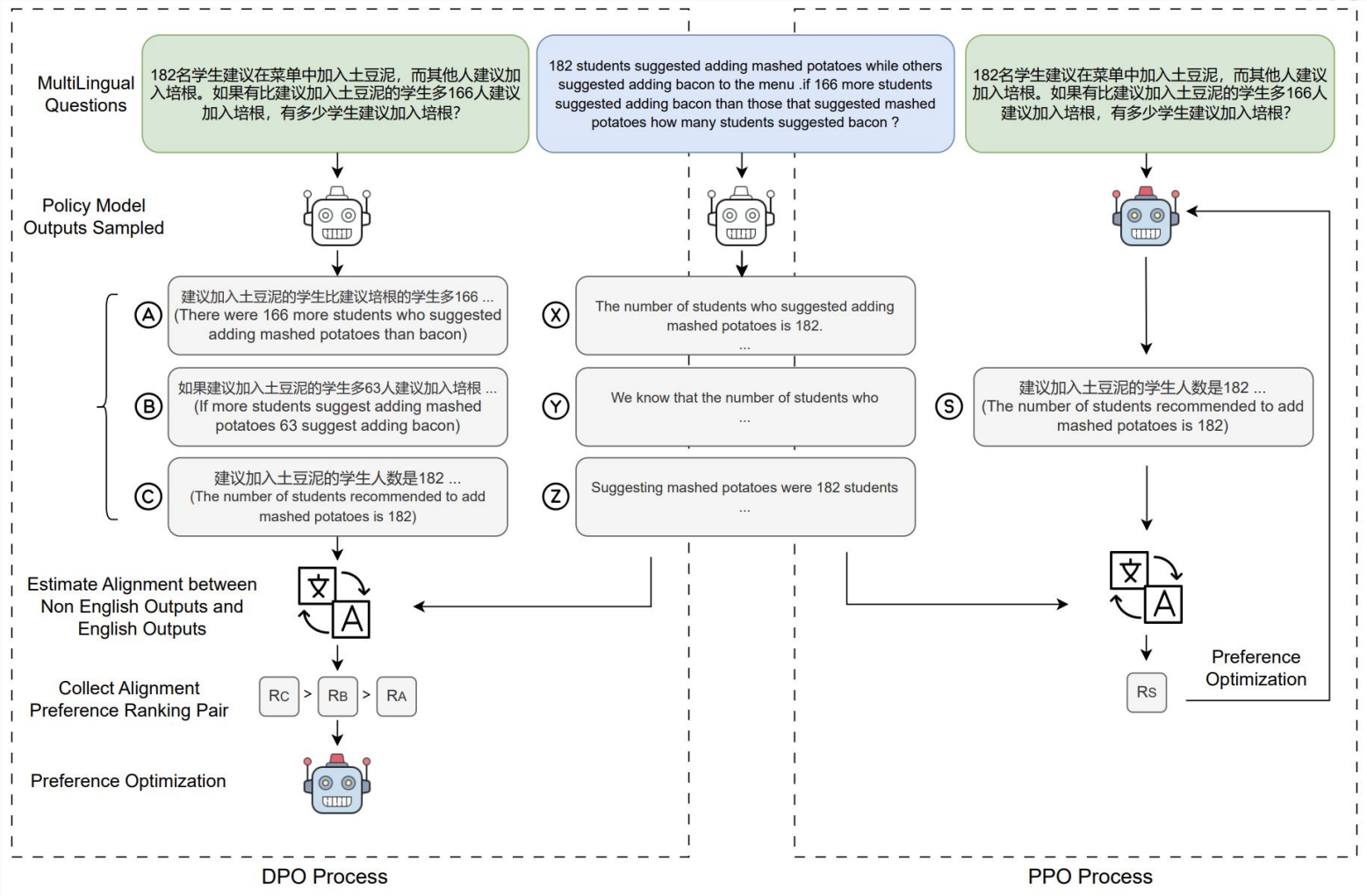
translate non-English model outputs into English language (worse)
translate English model outputs into non-English language (better) **+ DPO**



Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, Jiajun Zhang
"Language Imbalance Driven Rewarding for Multilingual Self-improving" (ICLR 2025)

MAPO: Multilingual-Alignment-as-Preference Optimization

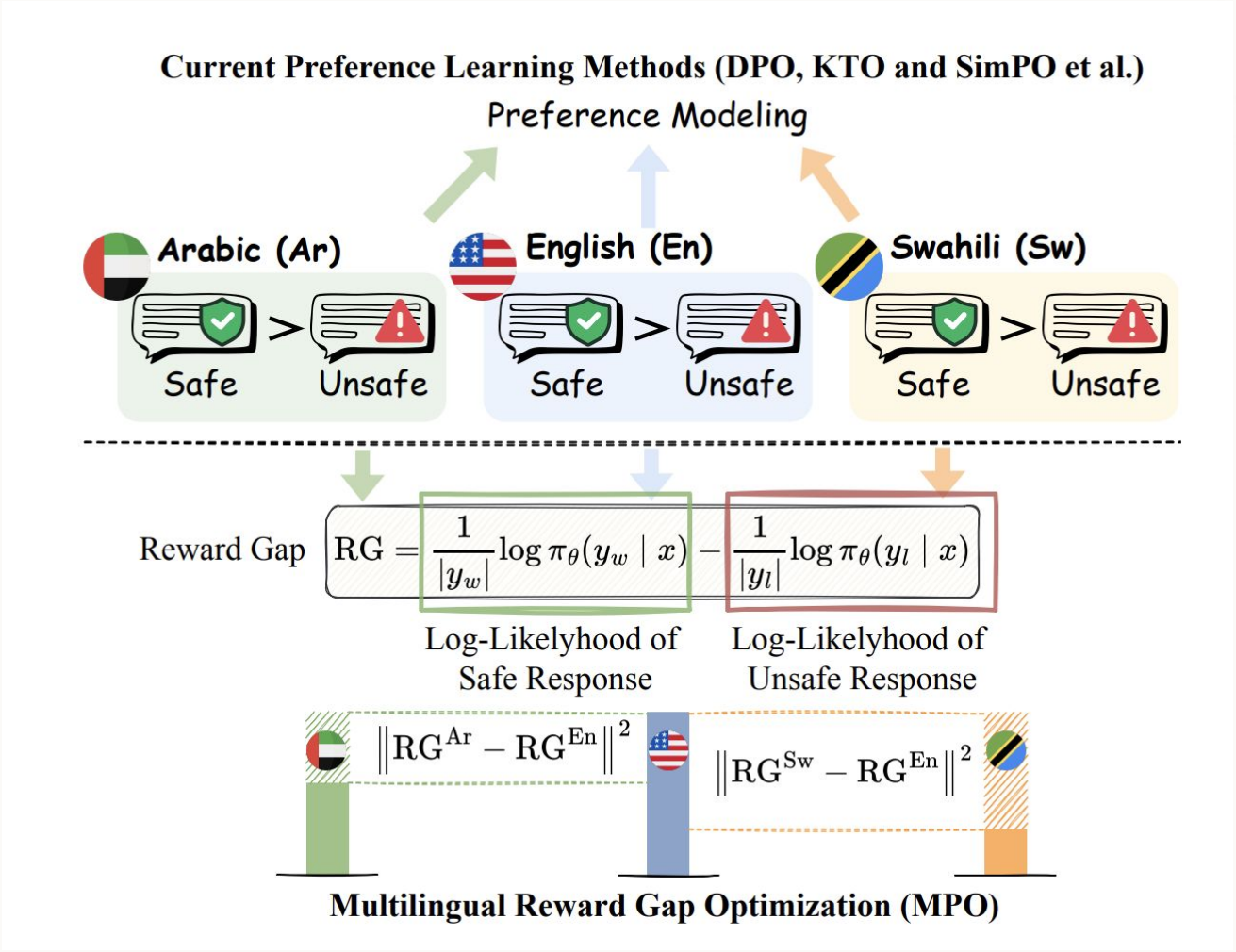
use a translation model to calculate the 'alignment' scores between answers in non-dominant and dominant languages as a preference



Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. "MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization" (ACL 2024)

MPO: Multilingual Reward GaP Optimization

Aligns the target language's reward gap (log-likelihood difference between preferred and dispreferred responses) with that of the dominant language (English).



Weixiang Zhao, Yulin Hu, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, Ting Liu
"MPO: Multilingual Safety Alignment via Reward Gap Optimization" (ACL 2026)

Multilingual Reinforcement Learning

⚔️ **Main Challenge**: multilingual rewards are hard to get right

Approaches	Advantages	Limitations
RLVR (Reinforcement Learning with Verifiable Rewards) [polyGRPO - Wu et al., 2026]	Language-agnostic; stable	Only works for verifiable tasks (e.g., math, coding) ☹
Pivot-Based Alignment [MAPO - She et al., 2024; MPO - Zhao et al., 2025]	Leverages strong English models; scalable	English-centric; may lose cultural/language nuances ☹
External, SFT-Based Reward Models [Nemotron-70B-Reward-Multilingual, 2025; mR3 - Anugraha et al., 2026]	Flexible; direct multilingual supervision ☺	require expensive, annotated data; risks poor RL performance when test tasks shift

Multilingual Reinforcement Learning

⚔️ **Main Challenge**: multilingual rewards are hard to get right

Approaches	Advantages	Limitations
RLVR (Reinforcement Learning with Verifiable Rewards) [polyGRPO - Wu et al., 2026]	Language-agnostic; stable	Only works for verifiable tasks (e.g., math, coding) ☹
Pivot-Based Alignment [MAPO - She et al., 2024; MPO - Zhao et al., 2025]	Leverages strong English models; scalable	English-centric; may lose cultural/language nuances ☹
External, SFT-Based Reward Models [Nemotron-70B-Reward-Multilingual, 2025; mR3 - Anugraha et al., 2026]	Flexible; direct multilingual supervision 🙂	require expensive, annotated data; risks poor RL performance when test tasks shift
Cross-Lingual Semantic Rewards [LRPO - Guo et al., 2026]	Flexible; direct multilingual supervision 🙂	Could be inconsistent across languages 💡 → but, can be calibrated to be more consistent!

Calibrate Multilingual Rewards

💡 **Calibrate cross-lingual semantic similarity to be more accurate rewards.**

Q: 竖大拇指👍在希腊🇬🇷通常是什么意思？

What does a 👍 thumbs-up usually mean in Greece 🇬🇷?

LLM-generated output $y^{(\ell_j)}$

في اليونان، عادةً ما يكون ذلك غير لائق وغير مقبول

(In Greece, this is usually considered inappropriate and unacceptable.)

human "gold" response $z^{(\ell_i)}$

在希腊带挑衅含义。

(It carries a provocative connotation in Greece.)

semantic similarity

- Qwen3-VL-Embedding [Li at al., 2026]
- mmBERT [Marone et al., 2025]
- NV-Embed [Lee et al., 2025]
- and many others …

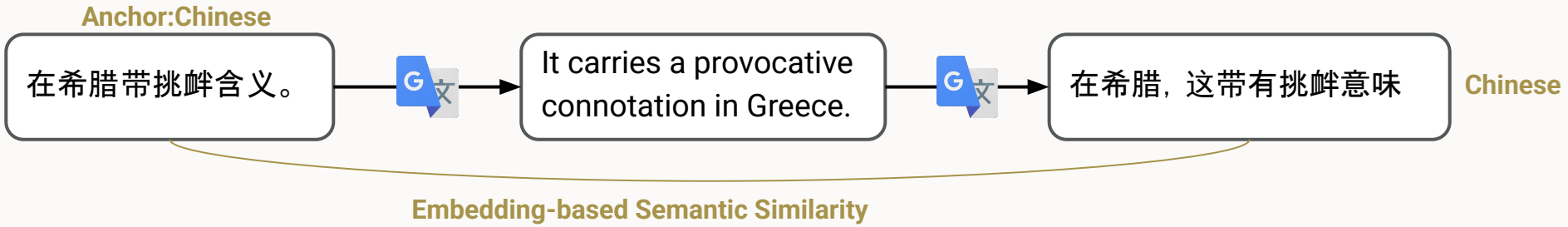
Reward Model ❄

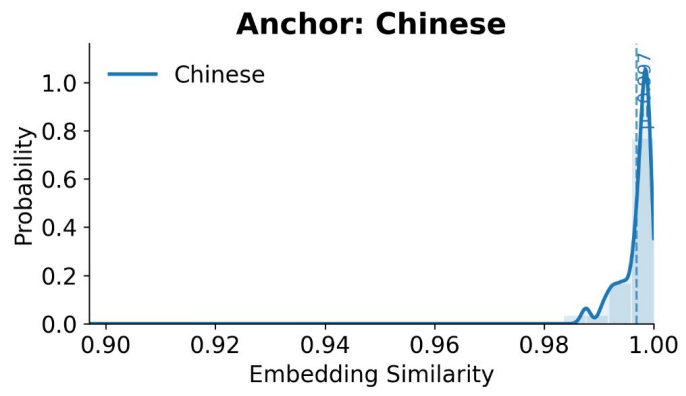
$$r^{\text{calib}}(y) = \mathcal{Q}_{\ell_i,\ell_j}\left(\text{sim}(y^{(\ell_j)}, z^{(\ell_i)})\right)$$

Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, Wei Xu
"Learning to Route Languages for Multilingual Policy Optimization" (ICML 2026)

Calibrate Multilingual Rewards

💡 **Calibrate cross-lingual semantic similarity to be more accurate rewards.**

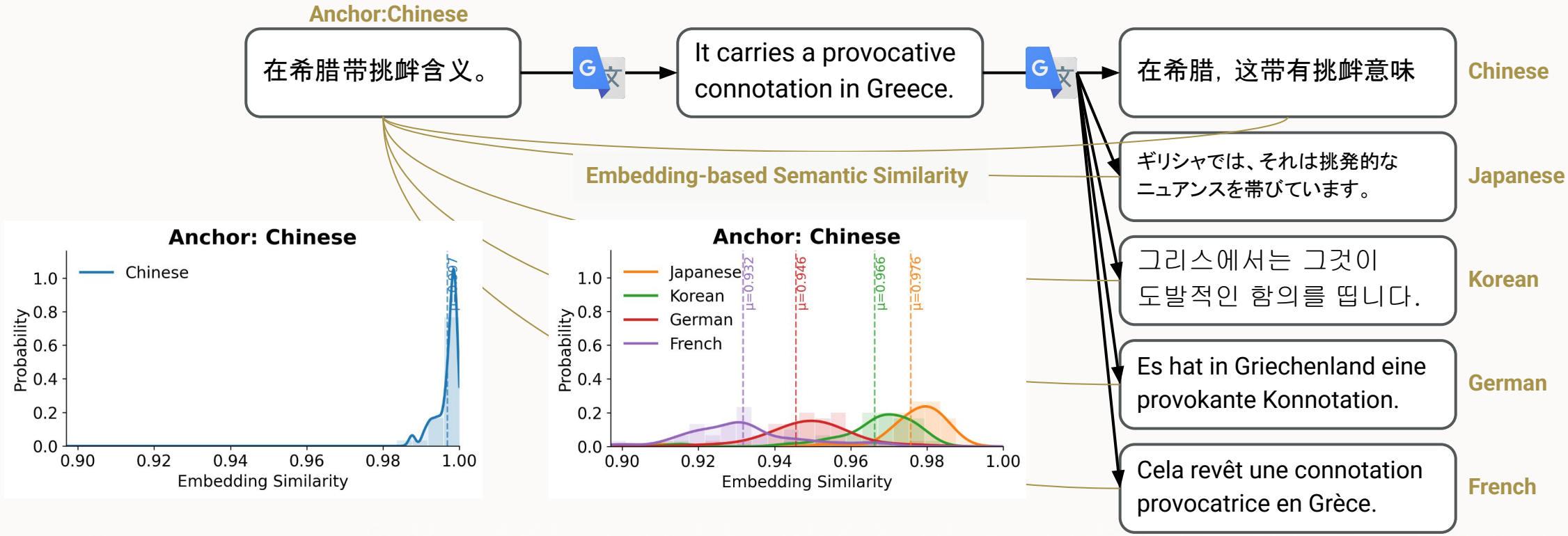




Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, Wei Xu
"Learning to Route Languages for Multilingual Policy Optimization" (ICML 2026)

Calibrate Multilingual Rewards

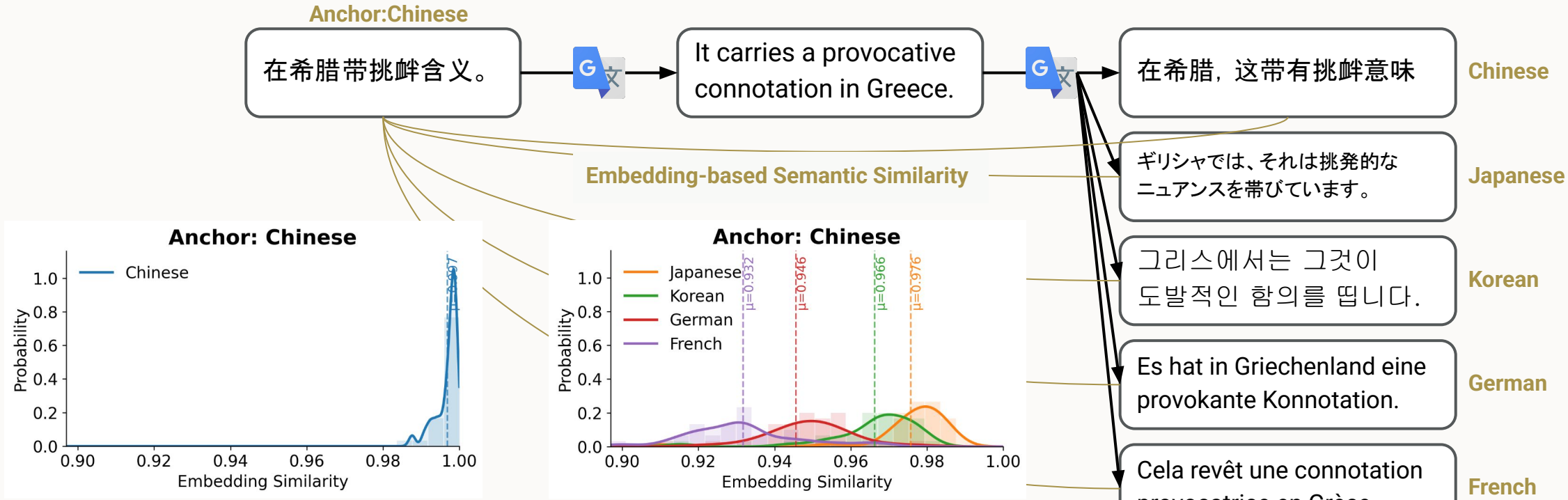
💡 **Calibrate cross-lingual semantic similarity to be more accurate rewards.**



Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, Wei Xu
"Learning to Route Languages for Multilingual Policy Optimization" (ICML 2026)

Calibrate Multilingual Rewards

💡 **Calibrate cross-lingual semantic similarity to be more accurate rewards.**



$$r^{\text{calib}}(y) = \text{sim}(y^{(\ell_j)}, z^{(\ell_i)}) - \lambda\left(\mu_{\ell_i,\ell_j} - \mu_{\text{ref}}\right)$$

Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, Wei Xu
"Learning to Route Languages for Multilingual Policy Optimization" (ICML 2026)

LRPO: Language-Routed Policy Optimization

💡 **Calibrate cross-lingual semantic similarity to be more accurate rewards.**

Reward Model ❄

Calibrated Reward
w/ language consistency

$$r = r^{\text{calib}} \cdot \mathbb{I}[\text{Lang}(y_k) = \ell_k]$$

$$\hat{A}_k = \frac{r_k - \text{mean}(r)}{\text{std}(r)}$$

Update with GRPO objective [Shao et al., 2025]

$$\mathcal{L}_{\text{GRPO}}(\pi_\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_k\}_{k=1}^{K}\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{K}\sum_{k=1}^{K} \left[\min\left(\frac{\pi_\theta(y_k|x)}{\pi_{\theta_{\text{old}}}(y_k|x)}\hat{A}_k, \text{clip}(\frac{\pi_\theta(y_k|x)}{\pi_{\theta_{\text{old}}}(y_k|x)}, 1-\epsilon, 1+\epsilon)\hat{A}_k \right) - \beta D_{\text{KL}}\left(\pi_\theta||\pi_{\text{ref}}\right) \right] \right]$$

Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, Wei Xu
"Learning to Route Languages for Multilingual Policy Optimization" (ICML 2026)

LRPO: Language-Routed Policy Optimization

LRPO = Dynamic Language Router + Calibrated Cross-lingual Rewards

Qwen2.5 | Regional Knowledge (5 languages) | Math Reasoning (10 languages) | Factual Knowledge (44 languages)

Approach	CARE	CARE (pro)			mGSM			Include-lite			Overall		
	Seen	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*
Vanilla	32.20	8.58	0.82	5.99	33.09	10.50	24.87	35.91	28.84	31.09	27.44	13.39	23.54

Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, Wei Xu
"Learning to Route Languages for Multilingual Policy Optimization" (ICML 2026)

LRPO: Language-Routed Policy Optimization

LRPO = Dynamic Language Router + Calibrated Cross-lingual Rewards

Qwen2.5 | Regional Knowledge (5 languages) | Math Reasoning (10 languages) | Factual Knowledge (44 languages)

Approach	CARE	CARE (pro)			mGSM			Include-lite			Overall		
	Seen	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*
Vanilla	32.20	8.58	0.82	5.99	33.09	10.50	24.87	35.91	28.84	31.09	27.44	13.39	23.54
MAPO	31.98	6.20	0.62	4.34	34.29	10.50	25.64	35.37	28.92	30.97	26.96	13.35	23.23
LIDR	30.07	6.44	0.62	4.50	32.17	8.60	23.60	35.06	28.66	30.69	25.93	12.63	22.21
MPO	31.91	6.58	1.03	4.73	32.97	11.20	25.05	35.62	29.10	31.17	26.77	13.78	23.22

We use the training split of two multilingual human preferences datasets, HelpSteer3 [Nvidia team, 2025] and CARE [Guo et al., 2025].

multilingual preference optimization by assuming a fixed dominant language (e.g., English) provides better supervision

Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, Wei Xu
"Learning to Route Languages for Multilingual Policy Optimization" (ICML 2026)

LRPO: Language-Routed Policy Optimization

LRPO = Dynamic Language Router + Calibrated Cross-lingual Rewards

Qwen2.5	Regional Knowledge (5 languages)			Math Reasoning (10 languages)			Factual Knowledge (44 languages)					
Approach	**CARE**	**CARE (pro)**			**mGSM**			**Include-lite**			**Overall**		
	Seen	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*
Vanilla	32.20	8.58	0.82	5.99	33.09	10.50	24.87	35.91	28.84	31.09	27.44	13.39	23.54
MAPO	31.98	6.20	0.62	4.34	34.29	10.50	25.64	35.37	28.92	30.97	26.96	13.35	23.23
LIDR	30.07	6.44	0.62	4.50	32.17	8.60	23.60	35.06	28.66	30.69	25.93	12.63	22.21
MPO	31.91	6.58	1.03	4.73	32.97	11.20	25.05	35.62	29.10	31.17	26.77	13.78	23.22
DPO	32.69	9.60	1.64	6.94	35.54	12.10	27.02	35.57	28.82	30.97	28.35	14.19	24.40

offline preference optimization using human-annotated preference data

Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, Wei Xu
"Learning to Route Languages for Multilingual Policy Optimization" (ICML 2026)

LRPO: Language-Routed Policy Optimization

LRPO = Dynamic Language Router + Calibrated Cross-lingual Rewards

Qwen2.5 | Regional Knowledge (5 languages) | Math Reasoning (10 languages) | Factual Knowledge (44 languages)

Approach	CARE	CARE (pro)			mGSM			Include-lite			Overall		
	Seen	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*
Vanilla	32.20	8.58	0.82	5.99	33.09	10.50	24.87	35.91	28.84	31.09	27.44	13.39	23.54
MAPO	31.98	6.20	0.62	4.34	34.29	10.50	25.64	35.37	28.92	30.97	26.96	13.35	23.23
LIDR	30.07	6.44	0.62	4.50	32.17	8.60	23.60	35.06	28.66	30.69	25.93	12.63	22.21
MPO	31.91	6.58	1.03	4.73	32.97	11.20	25.05	35.62	29.10	31.17	26.77	13.78	23.22
DPO	32.69	9.60	1.64	6.94	35.54	12.10	27.02	35.57	28.82	30.97	28.35	14.19	24.40

The fixed dominant language (i.e., English) cannot always provide better supervision training signal, especially on regional knowledge.

Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, Wei Xu
"Learning to Route Languages for Multilingual Policy Optimization" (ICML 2026)

LRPO: Language-Routed Policy Optimization

LRPO = Dynamic Language Router + Calibrated Cross-lingual Rewards

Qwen2.5	Regional Knowledge (5 languages)			Math Reasoning (10 languages)			Factual Knowledge (44 languages)					
Approach	**CARE**	**CARE (pro)**			**mGSM**			**Include-lite**			**Overall**		
	Seen	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*
Vanilla	32.20	8.58	0.82	5.99	33.09	10.50	24.87	35.91	28.84	31.09	27.44	13.39	23.54
MAPO	31.98	6.20	0.62	4.34	34.29	10.50	25.64	35.37	28.92	30.97	26.96	13.35	23.23
LIDR	30.07	6.44	0.62	4.50	32.17	8.60	23.60	35.06	28.66	30.69	25.93	12.63	22.21
MPO	31.91	6.58	1.03	4.73	32.97	11.20	25.05	35.62	29.10	31.17	26.77	13.78	23.22
DPO	32.69	9.60	1.64	6.94	35.54	12.10	27.02	35.57	28.82	30.97	28.35	14.19	24.40
GRPO	33.66	8.26	0.82	5.78	43.43	12.90	32.33	36.00	29.18	31.35	30.34	14.30	25.78

Online policy optimization method GRPO serves as a strong baseline.

Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, Wei Xu
"Learning to Route Languages for Multilingual Policy Optimization" (ICML 2026)

LRPO: Language-Routed Policy Optimization

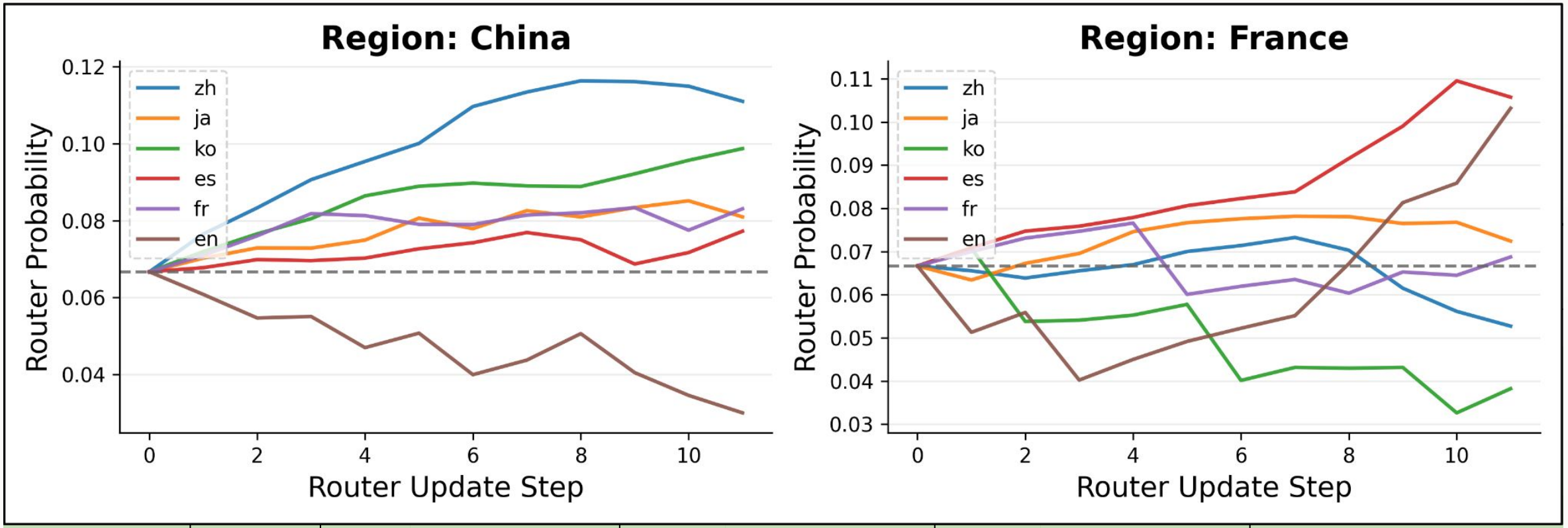
LRPO = Dynamic Language Router + Calibrated Cross-lingual Rewards

Qwen2.5	Regional Knowledge (5 languages)			Math Reasoning (10 languages)			Factual Knowledge (44 languages)					
Approach	**CARE**	**CARE (pro)**			**mGSM**			**Include-lite**			**Overall**		
	Seen	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*
Vanilla	32.20	8.58	0.82	5.99	33.09	10.50	24.87	35.91	28.84	31.09	27.44	13.39	23.54
MAPO	31.98	6.20	0.62	4.34	34.29	10.50	25.64	35.37	28.92	30.97	26.96	13.35	23.23
LIDR	30.07	6.44	0.62	4.50	32.17	8.60	23.60	35.06	28.66	30.69	25.93	12.63	22.21
MPO	31.91	6.58	1.03	4.73	32.97	11.20	25.05	35.62	29.10	31.17	26.77	13.78	23.22
DPO	32.69	9.60	1.64	6.94	35.54	12.10	27.02	35.57	28.82	30.97	28.35	14.19	24.40
GRPO	33.66	8.26	0.82	5.78	43.43	12.90	32.33	36.00	29.18	31.35	30.34	14.30	25.78
LRPO (this work)	**34.60**	**12.36**	1.23	**8.65**	**52.51**	**13.30**	**38.25**	**37.16**	28.90	**31.52**	**34.16**	**14.48**	**28.26**

LRPO outperforms existing policy optimization methods.

Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, Wei Xu
"Learning to Route Languages for Multilingual Policy Optimization" (ICML 2026)

LRPO: Language-Routed Policy Optimization



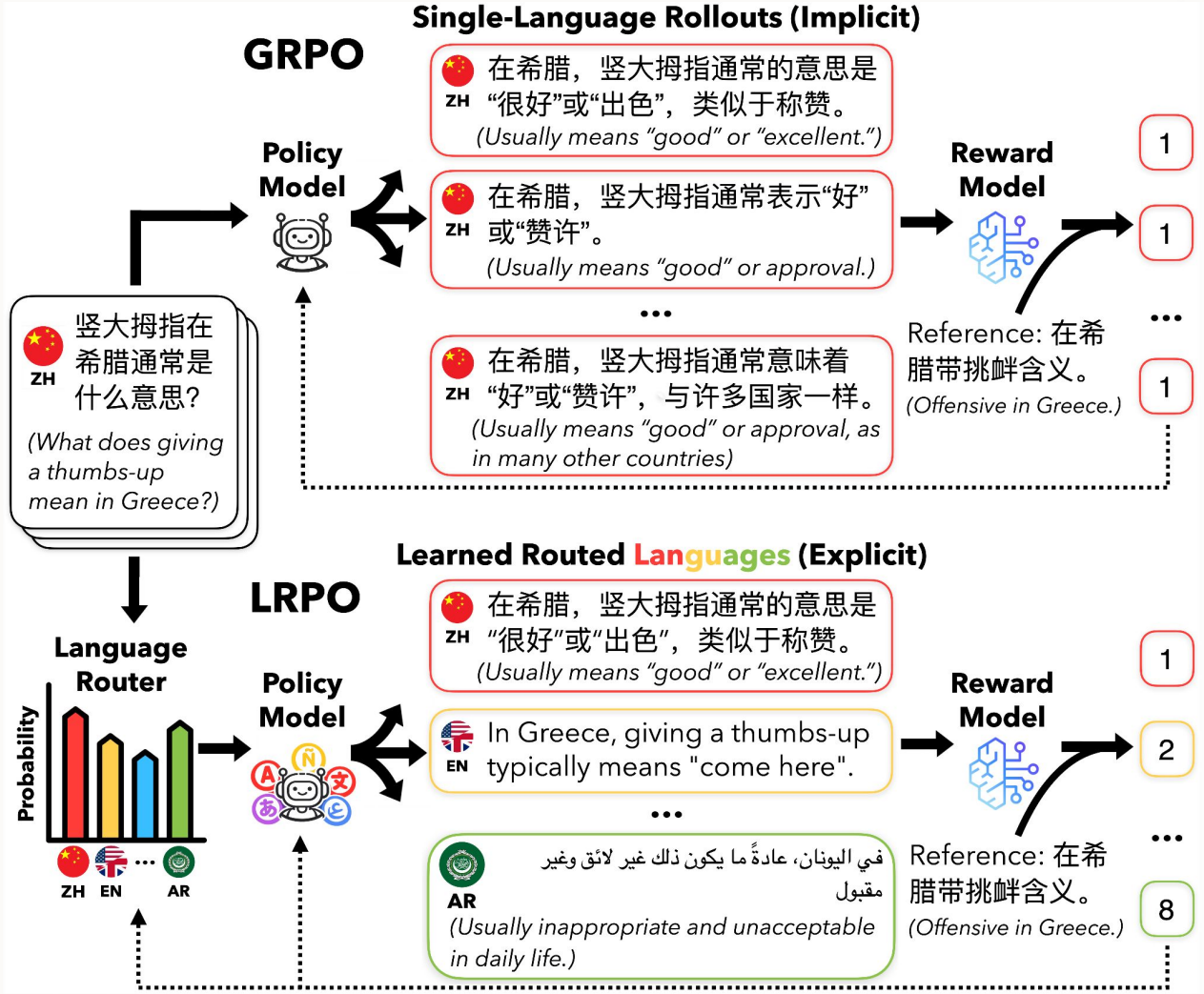
LRPO (this work)	**34.60**	**12.36**	1.23	**8.65**	**52.51**	**13.30**	**38.25**	**37.16**	28.90	**31.52**	**34.16**	**14.48**	**28.26**

Region-conditioned language probabilities evolve differently across regions.

Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, Wei Xu
"Learning to Route Languages for Multilingual Policy Optimization" (ICML 2026)

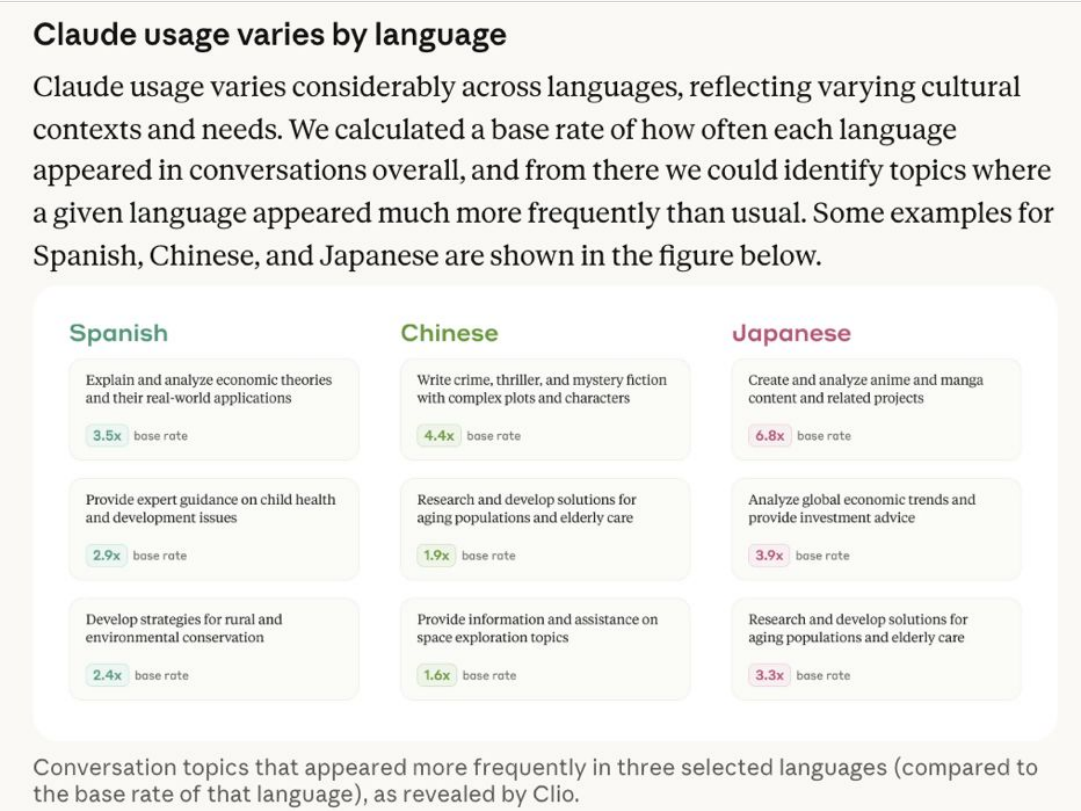
LRPO: Language-Routed Policy Optimization

Dynamic Language Router
(more controlled diversity in RL)
+
Calibrated Cross-lingual Rewards
(more flexible for any task)



Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, Wei Xu
"Learning to Route Languages for Multilingual Policy Optimization" (ICML 2026)

Our Findings Resonate with Others



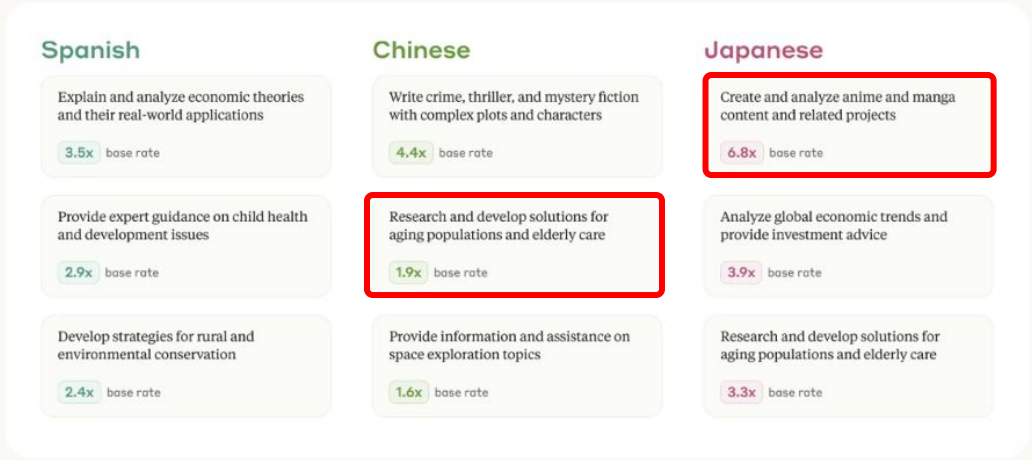
Topics also vary across languages

[Anthropic, Dec 2024]

Our Findings Resonate with Others

Claude usage varies by language

Claude usage varies considerably across languages, reflecting varying cultural contexts and needs. We calculated a base rate of how often each language appeared in conversations overall, and from there we could identify topics where a given language appeared much more frequently than usual. Some examples for Spanish, Chinese, and Japanese are shown in the figure below.

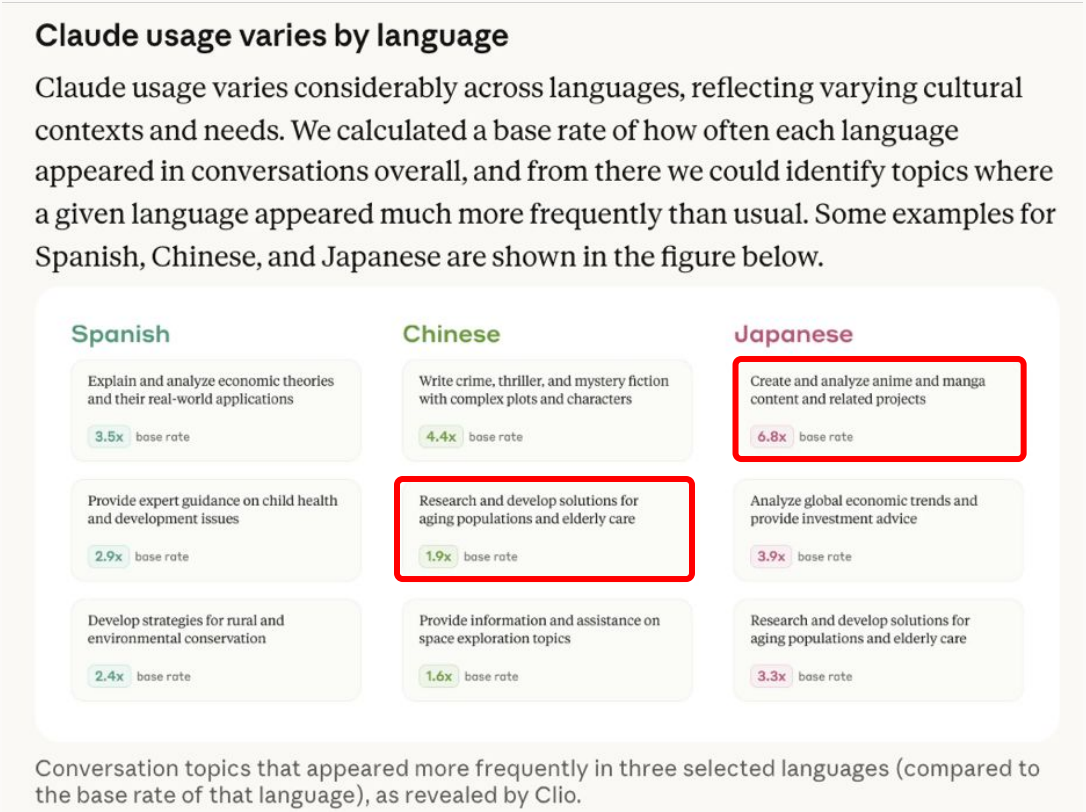


Conversation topics that appeared more frequently in three selected languages (compared to the base rate of that language), as revealed by Clio.

Topics also vary across languages

[Anthropic, Dec 2024]

Our Findings Resonate with Others



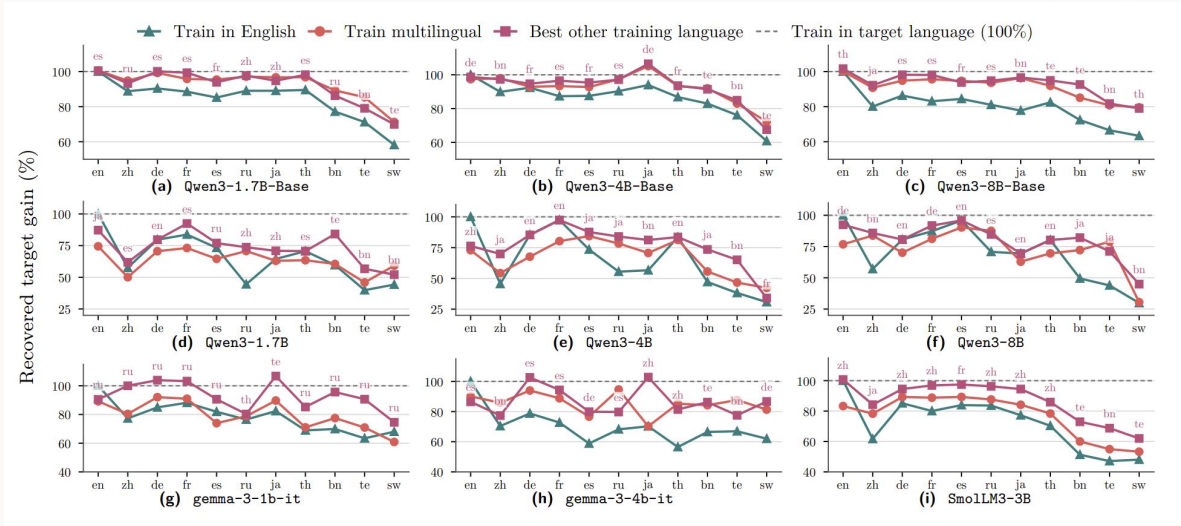
Claude usage varies by language

Claude usage varies considerably across languages, reflecting varying cultural contexts and needs. We calculated a base rate of how often each language appeared in conversations overall, and from there we could identify topics where a given language appeared much more frequently than usual. Some examples for Spanish, Chinese, and Japanese are shown in the figure below.

Conversation topics that appeared more frequently in three selected languages (compared to the base rate of that language), as revealed by Clio.

Topics also vary across languages

[Anthropic, Dec 2024]

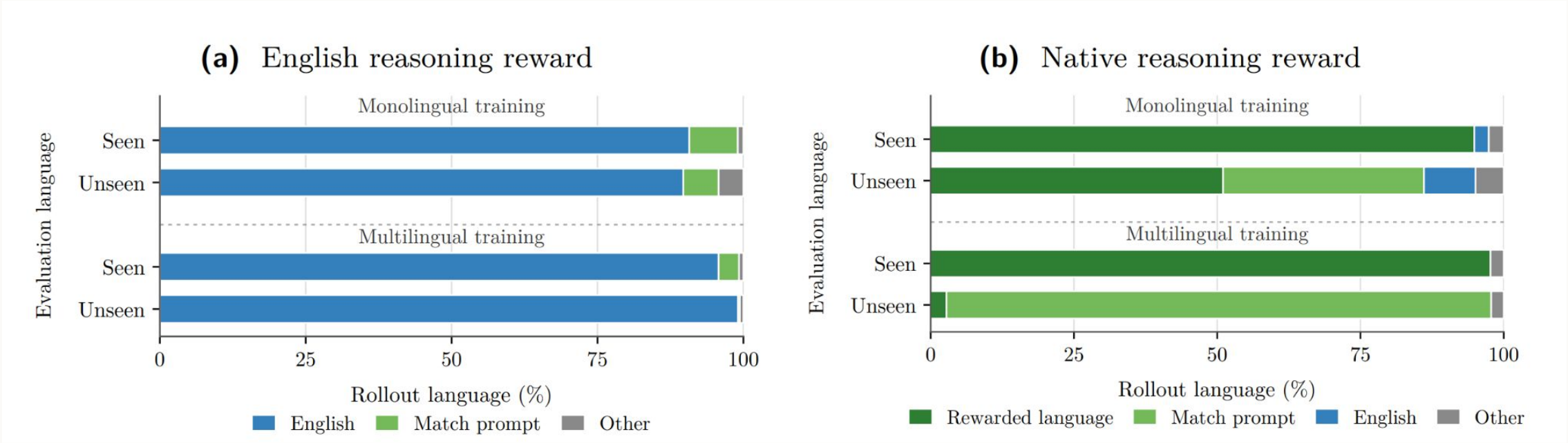


A large-scale Study of GRPO in multilingual settings

[Dobler et al., August 2026]

GRPO w/ Verifiable Rewards

Multilingual native-reward models consistently match the prompt language on both seen and unseen languages.

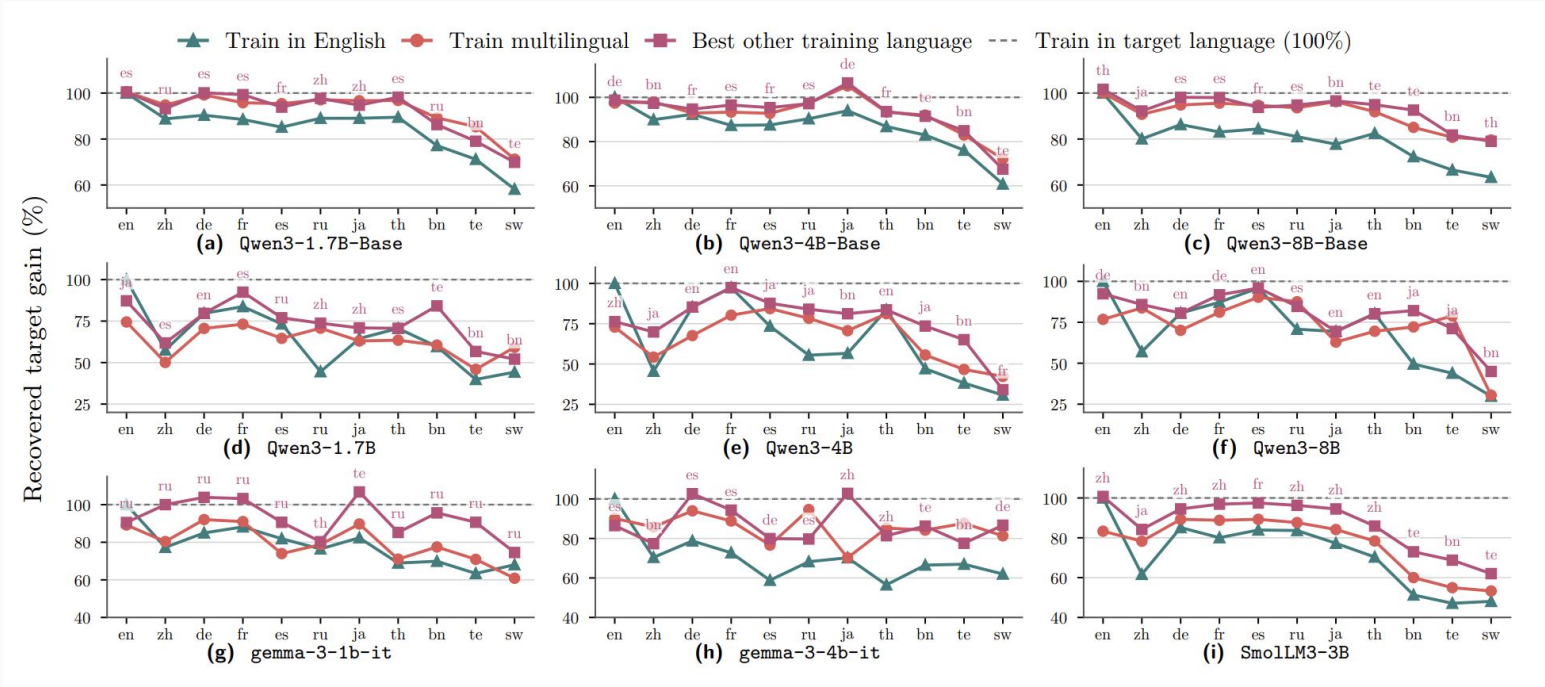


experiments on Multilingual Reasoning Gym [Shi et al., 2022] , that procedurally generated reasoning problems (e.g., algebra, arithmetic, logic, various games, etc.) across 14 languages, using hybrid human-machine translated templates

Konstantin Dobler, Federico Scozzafava, Jonathan Janke, Mohamed Ali, Simon Lehnerer
"GRPO Beyond English: A Large-Scale Study of GRPO in Non-English and Multilingual Settings"

GRPO w/ Verifiable Rewards

For in-domain tasks, training in the evaluation language (with monolingual rewards) often performs best, despite substantial cross-lingual transfer.



* Rewards: following standard RLVR, using a binary correctness reward when the extracted answer is verified as correct. Only when the answer is correct, use a combination of (1) a format reward for adhering to the specified output format and (2) a (English or Native) reasoning language reward which either rewards matching the prompt language or reasoning in English.

Konstantin Dobler, Federico Scozzafava, Jonathan Janke, Mohamed Ali, Simon Lehnerer
"GRPO Beyond English: A Large-Scale Study of GRPO in Non-English and Multilingual Settings"

Which leads us to ...

Learning from Multilingual Human Preference → **Multilingual Reinforcement Learning**

offline reinforcement learning → online reinforcement learning

Knowledge is not evenly distributed across languages.

Language/topic-aware routing and calibrated cross-lingual rewards are key to multilingual RL.

Which leads us to ...

Learning from Multilingual Human Preference → **Multilingual Reinforcement Learning** → **To Translate, or not to Translate — that's the Question**

offline reinforcement learning → online reinforcement learning

Knowledge is not evenly distributed across languages.

Language/topic-aware routing and calibrated cross-lingual rewards are key to multilingual RL.

Do you know what these words mean?

Doomscrolling

Barbiecore

chatfishing

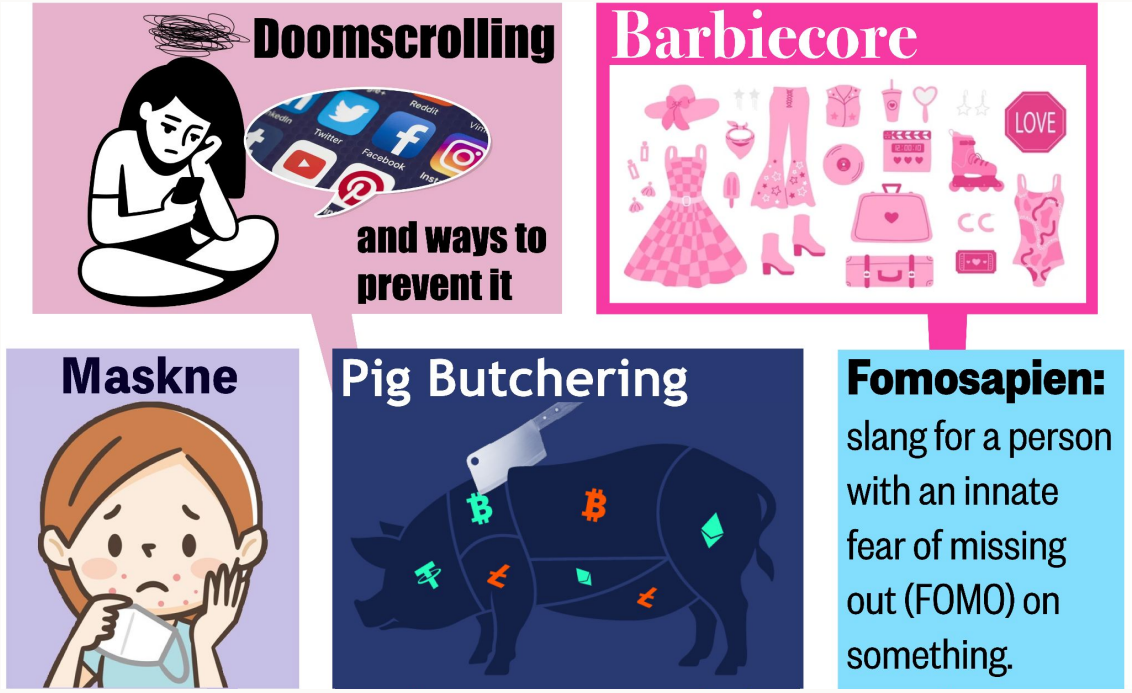
Maskne

Pig Butchering

Fomosapien

delulu with no solulu

Do you know what these words mean?



NEO-BENCH: Robustness of LLMs with Neologisms

[Zheng et al., 2024]

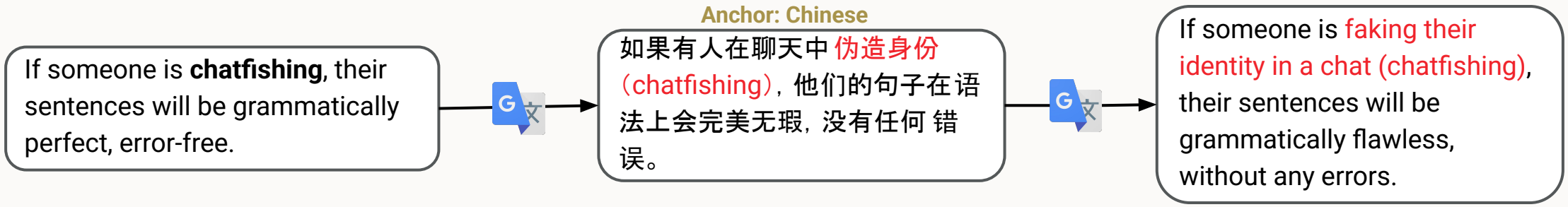
chatfishing

delulu with no solulu

Australia PM Anthony Albanese's speech

[March 2025]

Neologism is Lost in Translation



Neologism is Lost in Translation



[Zheng et al., EMNLP 2024] NEO-BENCH: Evaluating Robustness of Large Language Models with Neologisms

MGSM-Rev2: Multilingual Grade-school Math

English	How many girls are **not in** the girl scout?
German	Wie viele Mädchen sind **bei** den Pfadfinderinnen?
Backtranslation	How many girls are **in** the girl scout?
English	Artie has no change today, so he has decided to **round** all his prices to the nearest dollar.
German	Artie hat heute kein Wechselgeld, also entschließt er sich, alle Preise auf den nächsthöheren Dollarbetrag **aufzurunden**.
Backtranslation	Artie has no change today, so he has decided to **round up** all his prices to the nearest dollar.

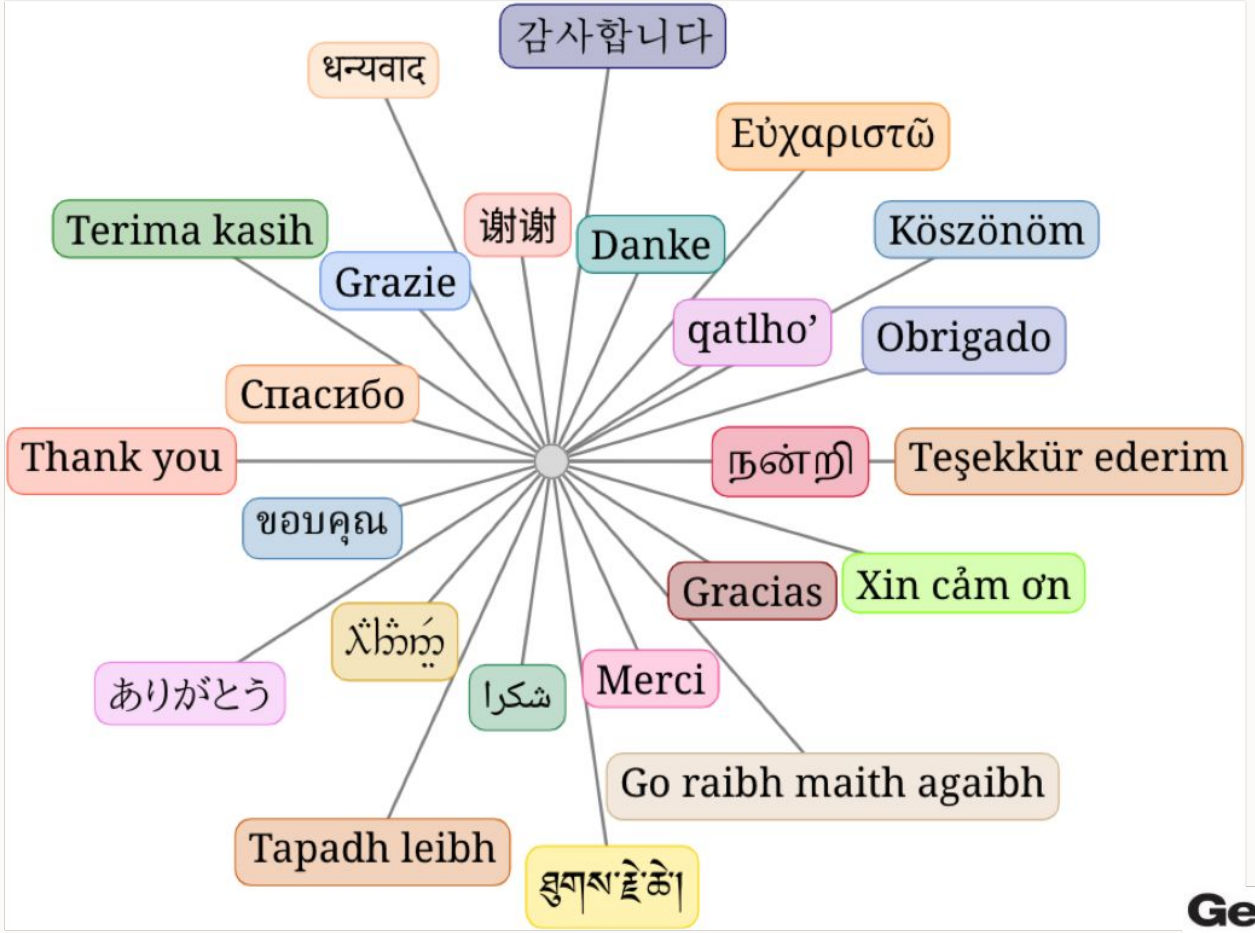
English	An orange costs **5 less** than what a watermelon cost.
German	Eine Orange kostet **5 Mal weniger** als eine Wassermelone.
Backtranslation	An orange costs **5 times less** than what a watermelon cost.
English	**Tuesday**, he walked 6 times as many miles as he walked on Monday.
German	Am **Donnerstag** ging er 6 Mal so viele Meilen, wie er am Montag gegangen war.
Backtranslation	**Thursday**, he walked 6 times as many miles as he walked on Monday.

a revised MGSM [Shi et al., 2022] benchmark w/ 250 grade-school math problems translated from English GSM8K [Cobbe et al., 2021] into 10 languages

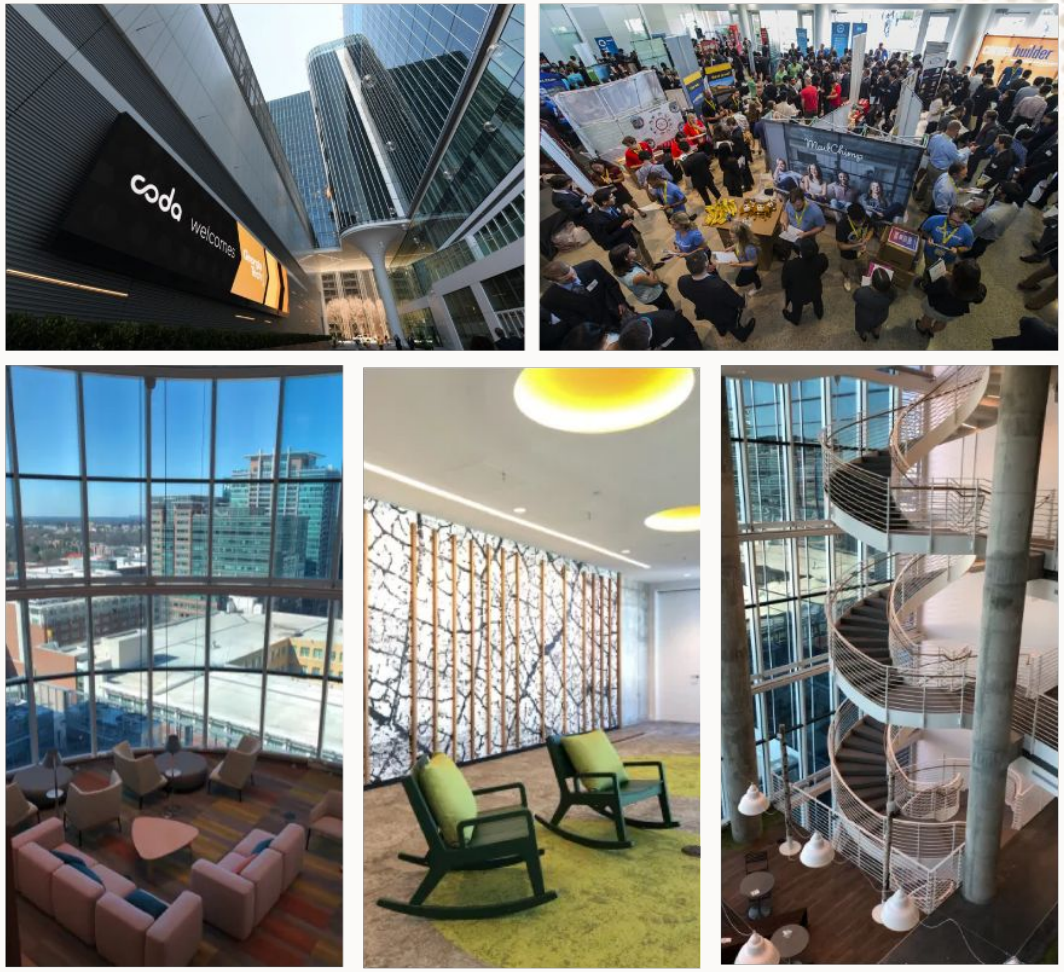
Jan-Thorsten Peter, David Vilar, Tobias Domhan, Dan Malkin, Markus Freitag. "Mind the Gap... or Not? How Translation Errors and Evaluation Details Skew Multilingual Results" (arXiv 2025)

Thank you!

https://cocoxu.github.io/



(image credit: Overleafl)



(image credit: Georgia Tech)

Georgia Tech Machine Learning

NSF

IARPA

DARPA

NIH

Google

SONY

criteo