M-GOAL: Toward Accurate and Language-Consistent Reasoning in LLMs via English-Only Post-Training

Dingyuan Dai[1,*], **Yuxi Chen**[2,*], **Jiayuan Wang**[1]
[1]University of California, Los Angeles, [2]University of Illinois Urbana-Champaign
ddai10@g.ucla.edu, yuxi5@illinois.edu, jiayuanwang@ucla.edu

Abstract

Reasoning-oriented language models inherit strong English-centric biases from pre-training, which are further reinforced by English-only post-training. This creates a dual failure mode where models fail both to reason correctly and to maintain reasoning traces in the query language. We study reasoning accuracy and language consistency under English-only post-training. SFT primarily improves language consistency with limited or inconsistent accuracy gains, while group relative policy optimization (GRPO) substantially improves accuracy at the cost of language consistency. Although translated multilingual GRPO improves both objectives, it requires translating reasoning post-training data into multiple languages and may introduce translation noise and information loss. We find that the language-consistency capability acquired by SFT and the accuracy capability acquired by GRPO are approximately orthogonal in the model's update space. By merging an English LoRA-SFT adapter with an English GRPO checkpoint trained from the same base model, M-GOAL improves MGSM accuracy from 39.1 to 57.8 and language consistency from 17.1 to 75.2 without translated reasoning supervision. This composition outperforms sequential training, its merge coefficient exposes a tunable operating point along the accuracy–consistency frontier, and the pattern holds beyond mathematical reasoning and on benchmarks whose non-English problems are natively authored rather than translated. Our results suggest that multilingual reasoning post-training does not necessarily require multilingual post-training data.

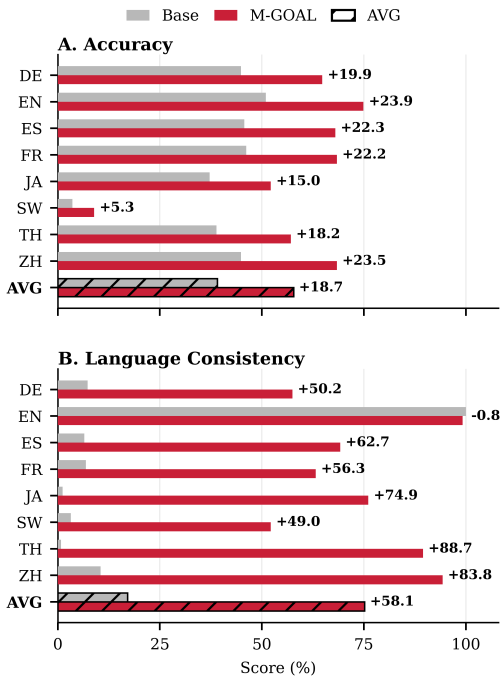


Figure 1: M-GOAL improves both reasoning accuracy and language consistency over the base model on MGSM using only English post-training data. Gray bars denote the Qwen3-1.7B base model, red bars denote M-GOAL, and numbers above the bars show absolute improvements in percentage points. M-GOAL improves the eight-language average from 39.1 to 57.8 accuracy and from 17.1 to 75.2 language consistency.

1 Introduction

Reasoning-oriented large language models produce an explicit chain of thought before committing to an answer (Wei et al., 2022), and reinforcement learning with verifiable rewards has become the dominant recipe for eliciting and sharpening this behavior (Shao et al., 2024; DeepSeek-AI et al., 2025). However, many widely used high-quality reasoning corpora are predominantly English (Cobbe et al., 2021; Hendrycks et al., 2021; Shao et al., 2024). This creates a mismatch with deployment because users often ask questions in other languages, while the model has been optimized largely on English reasoning.

In multilingual settings, this mismatch mani-

[*]Equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 26066–26084
October 24-29, 2026 ©2026 Association for Computational Linguistics

fests not only as reduced task accuracy (Shi et al., 2023), but also as a failure of language consistency, where the model reasons in a language other than the one the user used (Park et al., 2025; Barua et al., 2026). Recent multilingual CoT evaluations showed even under explicit target-language instructions, R1-Qwen-32B's sentence-level language compliance on MMMLU falls to 5% in French and Indonesian, 8% in Portuguese, and 15% in Korean (Zhao et al., 2026). Such failures are especially problematic when the intermediate chain of thought is user-facing, since they may shape how users trust or debug the model's answer (Wei et al., 2022; Turpin et al., 2023; Chen et al., 2025).

One approach is to introduce multilingual supervision by translating English reasoning data into target languages for training. Prior work shows that this strategy can be effective. Multilingual math instruction tuning improves cross-lingual reasoning performance (Chen et al., 2024), cross-lingual CoT tuning reduces performance gaps across languages (Chai et al., 2025), and fine-tuning on translated English reasoning traces outperforms training on target-language traces distilled from large reasoning models (Barua et al., 2026). But this relocates the problem to data construction: translated traces must be regenerated for every new target language, and their quality varies across languages and domains in ways that are hard to audit at scale (Shi et al., 2023; Xuan et al., 2025). This raises the question of whether both objectives can be improved without translated reasoning corpora.

In our study, we observe a consistent trade-off in English-only post training. Supervised fine tuning (SFT) improves language consistency but yields limited reasoning gains, while group relative policy optimization (GRPO) substantially improves reasoning accuracy at the cost of language consistency. We further find that language consistency is a sample-efficient and largely non-interfering adaptation: a small amount of direct non-English optimization signal suffices for GRPO to acquire target-language consistency while leaving its accuracy gains intact. This suggests that the two capabilities correspond to largely orthogonal updates in the model's update space, with language consistency emerging from a relatively lightweight adaptation. Motivated by this observation, we propose M-GOAL, short for Merged GRPO with an Orthogonal Adapter from LoRA-SFT. It starts from an English GRPO checkpoint and merges in a lightweight LoRA adapter learned by English SFT. As shown in Figure 1, this composition improves both reasoning accuracy and language consistency across evaluation languages, while using only English post-training data.

This paper makes three contributions. 1) We identify a trade-off between reasoning accuracy and language consistency in English-only SFT and GRPO post-training for multilingual reasoning, and show that the language-consistency capability acquired by SFT and the accuracy capability acquired by GRPO are approximately orthogonal in the model's update space, enabling their composition to obtain both capabilities simultaneously. 2) We show that mixed-language GRPO can resolve this trade-off but depends on translated data. 3) We propose a reliable parameter-merging method that combines English-only SFT and English-only GRPO, substantially improving both objectives without translated reasoning corpora, and show that its merge coefficient traces a controllable accuracy–consistency frontier.

2 Related Work

Prior work has shown that language models often lose reasoning accuracy outside English and may revert to English even when prompted in another language (Shi et al., 2023; Chen et al., 2024; Park et al., 2025; Barua et al., 2026). One line of work addresses this by adding target-language or multilingual supervision, including multilingual mathematical instruction tuning (Chen et al., 2024), cross-lingual CoT tuning (Chai et al., 2025), translated reasoning traces (Barua et al., 2026), and RL-based methods such as M-Thinker, which uses language-consistency and cross-lingual alignment rewards (Zhang et al., 2026).

A complementary line routes non-English reasoning through English: question-translation training aligns languages before unlocking English reasoning ability (Zhu et al., 2024), preference-based alignment derives cross-lingual consistency signals from an external translation model (She et al., 2024), TRIT integrates translation as a self-improving task within the RL loop (Liu et al., 2026), and representation-engineering approaches steer non-English hidden states toward English space at inference time (Li et al., 2025). These approaches all depend on target-language data, cross-lingual alignment objectives, or an English pivot;

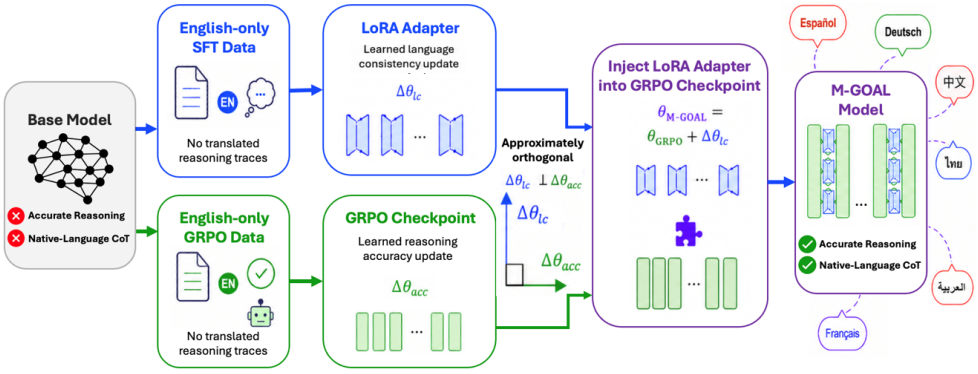


Figure 2: Overview of M-GOAL: an English-only SFT LoRA adapter learns a language-consistency update, while English-only GRPO learns a reasoning-accuracy update. M-GOAL composes these approximately orthogonal updates by injecting the LoRA adapter into the GRPO checkpoint, improving multilingual reasoning and native-language CoT without translated reasoning traces.

in contrast, our method reasons natively in the target language without routing through English or using translated reasoning supervision. That English-only tuning can elicit response-language matching at all is consistent with Shaham et al. (2024); we show it survives being detached into a LoRA adapter and re-attached to a separately trained RL policy.

3 Methods

Task and metrics. Given a question q in language $\ell(q)$, a model generates an output $o = (c, a)$, where c is the chain of thought and a is the final answer. For each example, we assume a gold answer $a^\star$ and evaluate the output along two axes:

$$\mathrm{Acc}(q, o) = 1[\mathrm{extract}(a) = a^\star], \quad (1)$$

$$\mathrm{LC}(q, o) = 1\left[p_{\ell(q)}(o) \geq \tau\right]. \quad (2)$$

Here, $\mathrm{extract}(a)$ reads the answer from the `\boxed{}` field, $p_{\ell(q)}(o)$ is the fraction of o detected as language $\ell(q)$ by a fastText classifier (Joulin et al., 2016), and $\tau = 0.5$. Our goal is to improve both answer accuracy and language consistency in non-English evaluation languages without using translated reasoning traces.

Models and data. We use Qwen3-1.7B-Base as our main model (Yang et al., 2025). To test whether our findings generalize across model families and scales, we additionally evaluate Qwen3-4B-Base (Yang et al., 2025) and SmolLM3-3B-Base (Bakouch et al., 2025). For mathematical reasoning, our English post-training data combines GSM8K (Cobbe et al., 2021) with a 12K-example subset of OpenR1 (Hugging Face, 2025). We filter the OpenR1 subset to remove image-dependent geometry problems, retain medium-length solutions, and select examples with moderate base-model difficulty. The final difficulty filter keeps examples with a Qwen3-1.7B-Base 8-rollout pass rate between $1/8$ and $6/8$. Full filtering details are provided in Appendix F.1.

We evaluate multilingual mathematical reasoning on MGSM (Shi et al., 2023), MSVAMP (Chen et al., 2024), and PolyMath-low (Wang et al., 2025). MGSM provides parallel grade-school math problems across languages and serves as our primary benchmark. MSVAMP provides a different multilingual arithmetic word-problem distribution, while PolyMath-low covers a broader multilingual math setting. Together, they test whether the accuracy–consistency trade-off holds across multiple multilingual math benchmarks.

For the non-mathematical ablation, we train on Explainable Commonsense Question Answering (ECQA) (Aggarwal et al., 2021) and evaluate cross-lingual transfer on MMMLU (OpenAI, 2024). This tests whether the SFT/GRPO split between language consistency and task accuracy extends beyond mathematical reasoning. Benchmark and dataset details are provided in Appendix D.

Training and evaluation. We compare supervised fine-tuning (SFT), group relative policy optimization (GRPO), sequential SFT→GRPO and GRPO→SFT baselines, and our parameter-merging method M-GOAL. SFT trains the model on English question–reasoning–answer demonstrations using standard token-level maximum

	German		English		Spanish		French		Japanese		Swahili		Thai		Chinese		Average	
Model	**Acc**	**LC**	**Acc**	**LC**	**Acc**	**LC**	**Acc**	**LC**	**Acc**	**LC**	**Acc**	**LC**	**Acc**	**LC**	**Acc**	**LC**	**Acc**	**LC**
Base	44.9	7.3	51.0	100	45.7	6.5	46.2	6.9	37.2	1.2	3.6	3.2	38.9	0.8	44.9	10.5	39.1	17.1
Full SFT	43.3	72.3	60.3	99.6	48.2	76.5	44.3	84.4	35.6	59.1	5.9	38.1	40.1	32.2	46.0	88.3	40.5	68.8
LoRA SFT	46.6	84.6	52.2	98.4	44.5	90.7	46.2	91.1	35.6	87.0	6.9	54.3	32.0	97.6	49.0	92.3	39.1	87.0
GRPO	70.9	0.8	87.0	100	74.9	0.0	71.7	0.0	55.9	0.0	7.3	0.8	64.8	0.0	70.9	0.0	62.9	12.7
SFT→GRPO	70.0	20.7	82.6	99.6	76.5	16.2	71.7	22.7	59.1	33.2	6.9	0.8	60.3	15.8	72.5	35.6	62.4	30.6
GRPO→SFT	48.6	74.5	60.3	100	53.0	70.9	51.4	82.2	43.7	78.5	5.7	11.3	40.9	63.2	55.9	89.5	44.9	71.3
M-GOAL	64.8	57.5	74.9	99.2	68.0	69.2	68.4	63.2	52.2	76.1	8.9	52.2	57.1	89.5	68.4	94.3	57.8	75.2

Table 1: MGSM results (%) for Qwen3-1.7B-Base under English-only post-training. We report answer accuracy (Acc) and language consistency (LC) for each evaluation language. M-GOAL is the only English-only method that improves both objectives strongly without translated reasoning supervision.

likelihood. GRPO trains the model on English questions with the standard clipped GRPO objective. The reward combines answer correctness and language consistency:

$$R(q, o) = 0.5\,\mathrm{Acc}(q, o) + 0.5\,\mathrm{LC}(q, o). \quad (3)$$

Since GRPO uses English-only training prompts, the LC term rewards matching English during training, rather than providing direct supervision for non-English language consistency.

M-GOAL is constructed by merging an English-trained LoRA-SFT adapter into an independently trained English GRPO checkpoint initialized from the same base model:

$$\theta_{\text{M-GOAL}} = \theta_{\text{GRPO}} + \Delta_{\text{LoRA-SFT}}. \quad (4)$$

We compare M-GOAL with the base model, the SFT-only and GRPO-only checkpoints, and the two sequential baselines, on MGSM, MSVAMP, and PolyMath-low, and on MMMLU after English-only ECQA post-training.

4 Results

4.1 The English-only landscape

Table 1 reports MGSM results for Qwen3-1.7B-Base under English-only post-training. The base model is weak on both objectives (39.1 Acc / 17.1 LC), and English-only post-training improves only one axis at a time: LoRA SFT raises LC to 87.0 while leaving accuracy unchanged, whereas GRPO raises accuracy to 62.9 while reducing LC to 12.7.

Sequential training trades the two objectives off but does not compose them: SFT→GRPO preserves accuracy (62.4) but recovers only modest consistency (30.6), while GRPO→SFT raises consistency (71.3) at the cost of accuracy (44.9). Among the English-only methods, M-GOAL achieves the strongest joint performance (57.8 / 75.2). Figure 3 visualizes the same pattern: SFT-based methods move along the consistency axis, GRPO-based methods along the accuracy axis, and only M-GOAL improves both. To verify that the high LC score reflects genuinely target-language reasoning rather than classifier artifacts, we manually inspected 100 randomly sampled M-GOAL outputs on MGSM with fastText target-language proportion above 0.5 (LC=1) and found no cases of substantive language mixing.

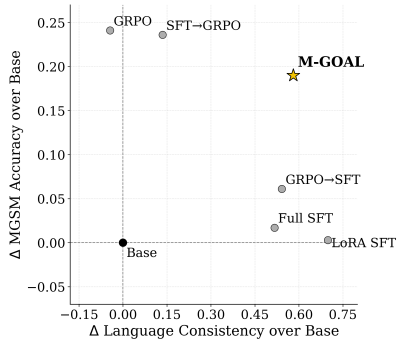


Figure 3: Improvements over the base model on MGSM (Qwen3-1.7B-Base), plotted as Δ language consistency (x-axis) against Δ accuracy (y-axis). GRPO-based methods gain accuracy but little or negative consistency (upper left); SFT methods gain consistency but almost no accuracy (lower right). M-GOAL (Gold Star) achieves the strongest balanced improvement across both axes.

M-GOAL does not match English-only GRPO in absolute accuracy: on MGSM it trades 5.1 accuracy points for a 62.5-point consistency gain, and GRPO likewise leads on accuracy on MSVAMP (73.9 vs. 67.3) and PolyMath-low (62.5 vs. 56.8) while scoring near zero on LC. This is an operating point rather than a fixed cost: Section 5.4 shows that the merge coefficient traces the whole accuracy–consistency frontier.

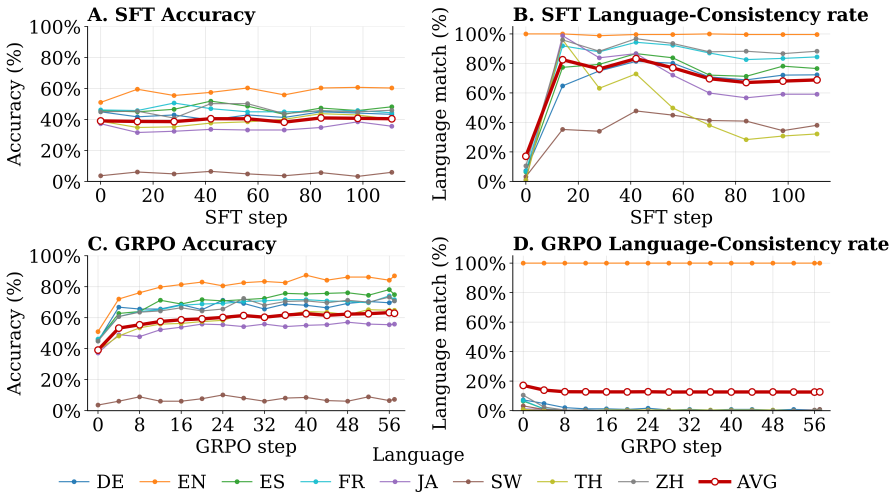


Figure 4: Training dynamics of English-only SFT and GRPO on MGSM using Qwen3-1.7B. Top row: SFT. Bottom row: GRPO.

	MSVAMP		PolyMath-low	
Model	Acc	LC	Acc	LC
Base	44.3	20.5	37.5	17.1
Full SFT	53.6	66.9	39.1	68.4
LoRA SFT	56.1	87.0	40.4	88.6
GRPO	73.9	12.7	62.5	12.8
SFT→GRPO	72.9	34.8	60.1	30.0
GRPO→SFT	54.1	73.6	42.9	72.9
M-GOAL	67.3	72.8	56.8	71.3

Table 2: Results on MSVAMP and PolyMath-low under English-only post-training. M-GOAL consistently improves both reasoning accuracy and language consistency across both benchmarks.

Across math benchmarks. Table 2 repeats the comparison on MSVAMP and PolyMath-low. The same trade-off appears on both benchmarks. GRPO gives the highest accuracy but the weakest language consistency, while SFT gives strong language consistency with limited accuracy gains. As on MGSM, M-GOAL is the only compared variant that maintains both high accuracy and high language consistency (67.3/72.8 on MSVAMP and 56.8/71.3 on PolyMath-low).

Model	Accuracy	Language Consistency
Base	32.3	11.9
Full SFT	32.4	75.5
LoRA SFT	34.9	65.3
GRPO	51.3	13.4
M-GOAL	45.3	63.9

Table 3: Ablation beyond mathematical reasoning on MMMLU. M-GOAL improves both accuracy and language consistency under English-only post-training.

	C-MGSM		CM17K-zh		Ape210K-zh		DMath-ko	
Model	Acc	LC	Acc	LC	Acc	LC	Acc	LC
Base	44.0	19.2	35.7	12.0	42.5	2.6	31.5	3.5
LoRA SFT	52.0	76.3	42.0	68.7	49.8	40.4	38.1	66.0
GRPO	71.0	14.5	74.7	0.0	75.2	0.0	57.9	2.0
M-GOAL	67.1	65.6	70.2	85.4	71.0	54.1	55.6	59.5

Table 4: Benchmarks whose non-English problems are natively authored rather than translated from English. C-MGSM numbers are seven-language averages; the other three are monolingual Chinese or Korean splits.

Natively authored, non-translated benchmarks. MGSM, MSVAMP, and PolyMath-low all obtain their non-English splits by translating source problems, so their items may carry translationese. We therefore also evaluate four benchmarks with natively authored non-English problems: C-MGSM, a Chinese-source counterpart of MGSM that we build from natively written Chinese elementary-school problems and then translate into the other evaluation languages; and the monolingual Chinese splits of CM17K (Qin et al., 2021) and Ape210K (Zhao et al., 2020) together with the Korean split of DMath (Kim et al., 2023). Table 4 shows the same structure as before, with M-GOAL again retaining most of the GRPO accuracy gain while restoring consistency (Appendix D).

Beyond math and across model families. The same pattern also holds beyond mathematical reasoning and across model families. On MMMLU after English-only ECQA post-training (Table 3), GRPO achieves the highest accuracy but low

language consistency, SFT improves consistency with little accuracy gain, and M-GOAL improves both axes over the base model. On additional base models (Table 5), M-GOAL again improves both objectives: Qwen3-4B-Base improves from 56.4 to 69.8 Acc and from 20.2 to 82.5 LC, while SmolLM3-3B-Base improves from 18.2 to 54.3 Acc and from 17.0 to 44.6 LC. The smaller consistency gain on SmolLM3-3B-Base is consistent with a stronger English-centric pretraining bias. The trade-off and the benefit of parameter merging are thus not specific to MGSM or Qwen3-1.7B.

Model	MGSM Acc	MGSM LC
Qwen3-4B-Base		
Base	56.4	20.2
Full SFT	57.8	73.6
LoRA SFT	58.7	85.8
GRPO	78.3	12.8
M-GOAL	69.8	82.5
SmolLM3-3B-Base		
Base	18.2	17.0
Full SFT	32.2	53.5
LoRA SFT	41.6	51.6
GRPO	67.3	14.9
M-GOAL	54.3	44.6

Table 5: MGSM results across additional base models under English-only post-training. M-GOAL improves both answer accuracy and language consistency over the corresponding base model without multilingual post-training data.

5 Analysis

5.1 How English-only Post-training Shapes Accuracy and Language Consistency

English-only SFT and GRPO learn complementary abilities. Figure 4 compares the training dynamics of English-only SFT and GRPO. SFT rapidly improves language consistency within the first few steps, while accuracy remains nearly flat. This suggests that matching the reasoning language to the question language is a relatively lightweight adaptation. GRPO shows the opposite pattern: accuracy increases steadily, but language consistency on multilingual evaluation quickly collapses. Since the GRPO training prompts are English-only, the LC reward provides no direct optimization signal for non-English consistency.

The collapse is not an artifact of the reward. Equation 3 contains an LC term that reduces to rewarding English under English-only prompts, so in principle the collapse could be built into the reward. An accuracy-only control trained with identical data and hyperparameters gives 63.2 Acc / 12.6 LC against 62.9 / 12.7 for the mixed reward, i.e. no difference on either axis. The reason is that with English prompts nearly every rollout in a group is English, so the binary LC term is constant within the group and cancels in the group-normalized advantage. The collapse comes from the English-only training distribution rather than the reward design (Appendix G).

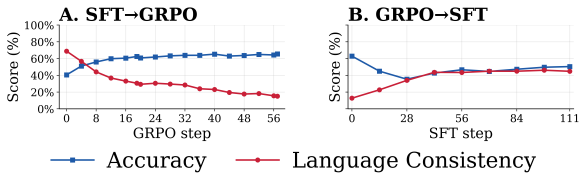


Figure 5: Sequential post-training fails to preserve both capabilities under English-only data.

Sequential training overwrites the previous capability. Figure 5 shows that continuing optimization in either order fails to preserve both abilities. In SFT→GRPO, the language consistency acquired during SFT is rapidly degraded, while the resulting accuracy gain is no better than direct GRPO from the base model. In GRPO→SFT, SFT recovers some language consistency but substantially reduces the accuracy gained during GRPO, and the recovered consistency remains below direct SFT from the base model.

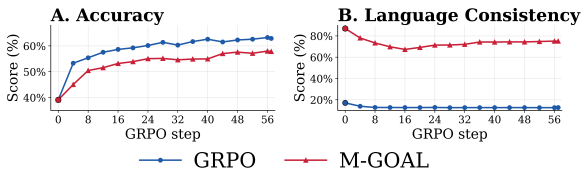


Figure 6: Merging LoRA-SFT into GRPO checkpoints across training. The merged model preserves most of the GRPO accuracy trajectory while maintaining much higher language consistency than GRPO alone at every step.

Parameter merging avoids sequential interference. In contrast to sequential training, parameter merging preserves both capabilities more effectively. Figure 6 compares the GRPO trajectory with and without merging the LoRA-SFT adapter. The merged model largely follows the accuracy trajectory of GRPO, with only a small accuracy gap, while maintaining substantially higher

	Accuracy	Language Consistency
Base	39.1	17.1
SFT_{en}	39.1	87.0
$SFT_{translated}$	37.7	78.2

Table 6: Effect of translating SFT data on MGSM. SFT_{en} uses English-only LoRA SFT; $SFT_{translated}$ adds translated data.

language consistency at every GRPO step. This indicates that the LoRA-SFT adapter provides a reusable language-consistency direction that can be composed with GRPO checkpoints without overwriting the accuracy capability.

5.2 Where translated data helps

We next ask whether adding translated training data lets each method improve on its weaker objective. We translate our English-only dataset into Chinese, Spanish, and Thai using DeepSeek-V4-Flash (DeepSeek-AI, 2026), verify the language of every translated question automatically with fastText, and manually inspect a subset for quality.

Translating SFT data does not help. Since SFT already achieves strong language consistency, we test whether adding Chinese and Spanish translations to the English SFT data brings further gains. Table 6 shows that it leaves accuracy essentially unchanged and slightly lowers language consistency.

Training data	Acc	LC
0% Trans. (EN-only)	62.9	12.7
50% EN + 50% ES	63.9	91.2
50% EN + 50% ZH	64.7	84.6
50% EN + 50% TH	68.1	89.2

Table 7: Effect of balanced translated GRPO data on MGSM for Qwen3-1.7B.

Translated GRPO data resolves the trade-off. We next test whether translated prompts can provide the missing language consistency signal during GRPO. For each target language, we train a balanced mixture with 50% English prompts and 50% translated prompts (Figure 7; Table 7). Across Spanish, Chinese, and Thai, translated GRPO keeps accuracy comparable to English-only GRPO while raising LC from 12.7 to 91.2, 84.6, and 89.2, respectively. LC rises sharply in the first few GRPO steps and then saturates, while accuracy continues to improve. This suggests that language consistency is a sample-efficient adaptation under GRPO, with little interference to accuracy gains. However, translated GRPO requires target-language post-training prompts and may introduce translation noise or information loss, motivating our English-only setting. Additional translated-GRPO ratio experiments are provided in Appendix C.

What the English-only constraint buys. Translated GRPO requires, for every new target language, a translated prompt corpus, a quality-control pass over it, and its own training run; one English-only M-GOAL model serves every evaluated language. Translation is not uniformly reliable either: about 30% of our Thai prompts failed automatic checks and needed retranslation, best-of-n resampling with an NLLB fallback, or manual repair. The mixture ratio is a further hyperparameter that does not transfer across tasks: on mathematical reasoning a 10% translated fraction is already near saturation, whereas on ECQA→MMMLU it raises Spanish LC only from 13.4 to 14.3, and the mixture must approach 50% before LC comes near the 74.6 that M-GOAL reaches on that split with no translated data (Appendix C.1). Translated GRPO remains attractive when high-quality target-language prompts already exist; M-GOAL targets the case where they do not.

5.3 Mechanism Analysis: Why Can LoRA-SFT and GRPO Be Composed?

The previous results show that merging an SFT-trained LoRA adapter into a GRPO checkpoint preserves most of both capabilities, whereas sequential training in either order overwrites one of them. We hypothesize that this difference arises because the language-consistency capability acquired by SFT and the accuracy capability acquired by GRPO occupy approximately orthogonal directions in the model's update space. Under this hypothesis, parameter addition can compose the two updates without destructive interference, following the task-vector view of model editing (Ilharco et al., 2023).

Specificity: the LoRA gain is learned, not generic capacity. One possible explanation is that any same-rank LoRA perturbation could help by adding capacity. We test this with a random-LoRA control. The random adapter has the same rank, target modules, and per-layer Frobenius

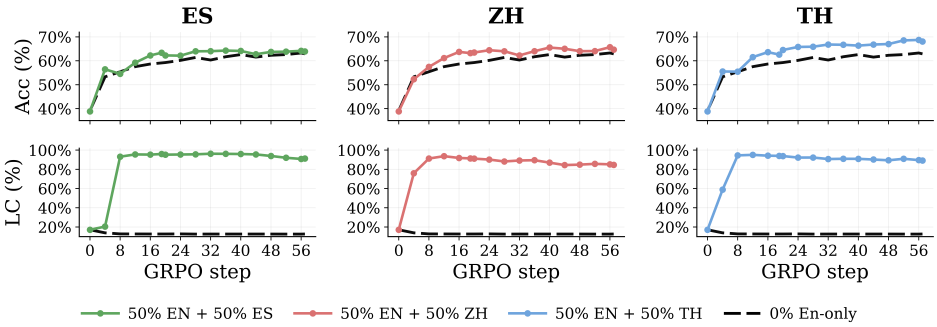


Figure 7: GRPO training dynamics with balanced English–translated mixtures on MGSM. Each run uses 50% English prompts and 50% prompts translated into one target language: Spanish, Chinese, or Thai. Top row: accuracy; bottom row: language consistency. The dashed black curve denotes English-only GRPO. Translated GRPO raises language consistency early while preserving the accuracy trajectory.

norms as the LoRA-SFT adapter, but is randomly initialized and never trained.

Configuration	ACC	LC
No-LoRA (GRPO only)	62.9	12.7
+ A1 (learned LoRA-SFT)	57.8	75.2
+ A1-randinit (random LoRA)	63.2	12.7

Table 8: Random-LoRA control. A matched random LoRA gives no LC gain, while the learned LoRA-SFT restores LC.

As shown in Table 8, the random LoRA leaves LC unchanged at 0.127. In contrast, the learned LoRA raises LC to 0.752. Accuracy is similar for GRPO and random LoRA. Thus, the LC gain comes from the SFT-learned direction, not from added low-rank capacity.

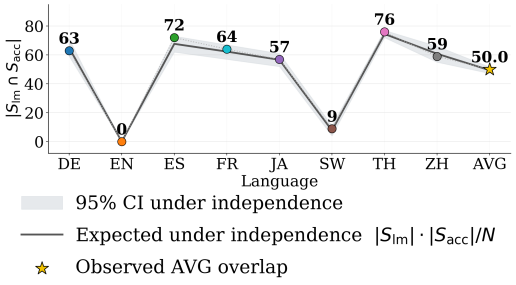


Figure 8: Behavioral independence between LoRA-SFT language-consistency gains and GRPO accuracy gains on MGSM. For each language, we compare the observed joint-improvement count with the count expected under independence, $|S_{\mathrm{LC}}||S_{\mathrm{Acc}}|/N$. Error bars denote the 95% confidence interval under the independence null.

Behavioral independence: the two gains affect different samples. We next test whether the language-consistency gain from LoRA-SFT and the accuracy gain from GRPO occur on the same examples. For each MGSM example x_i, we define two improvement sets: S_{LC}, where LoRA-SFT improves language consistency over the base model, and S_{Acc}, where GRPO improves answer accuracy over the base model. If the two gains are independent, the expected overlap is

$$\mathbb{E}\left[|S_{\mathrm{LC}} \cap S_{\mathrm{Acc}}|\right] = \frac{|S_{\mathrm{LC}}||S_{\mathrm{Acc}}|}{N}.$$

Figure 8 shows that the observed overlap matches this independence expectation for each language. All observed counts fall within the 95% confidence interval of the independence null. In the pooled evaluation set, the observed overlap is 400, close to the expected 395.6. Thus, LoRA-SFT and GRPO improve nearly independent subsets of examples, supporting the view that they act on separate behavioral directions.

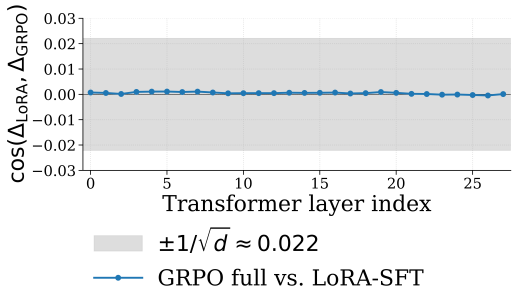


Figure 9: Per-layer cosine similarity between the LoRA-SFT update and the GRPO update. The shaded band marks the expected random-vector scale, $\pm 1/\sqrt{d}$, where d is the dimensionality of each flattened layer update.

Weight-space orthogonality: the two updates point in different directions. Finally, we directly measure the geometry of the two updates

in weight space, following the task-vector view (Ilharco et al., 2023). For each LoRA-targeted layer W, we flatten the LoRA-SFT update and the GRPO full-parameter update, and compute their cosine similarity:

$$\cos(W) = \frac{\left\langle \mathrm{vec}(\Delta^{(W)}_{\mathrm{LoRA-SFT}}), \mathrm{vec}(\Delta^{(W)}_{\mathrm{GRPO}}) \right\rangle}{\left\| \Delta^{(W)}_{\mathrm{LoRA-SFT}} \right\|_F \left\| \Delta^{(W)}_{\mathrm{GRPO}} \right\|_F},$$

where $\Delta^{(W)}_{\mathrm{LoRA-SFT}}$ is the LoRA-SFT update at layer W and $\Delta^{(W)}_{\mathrm{GRPO}} = \theta^{(W)}_{\mathrm{GRPO}} - \theta^{(W)}_{\mathrm{base}}$ is the corresponding GRPO full-parameter update. Figure 9 shows that the per-layer cosine remains near zero across all transformer layers and stays within the random-vector scale. We also compute a whole-subspace cosine by concatenating all LoRA-targeted layer updates before taking the cosine. This value is 0.0004, confirming that the near-orthogonality also holds for the combined update direction. These results indicate that the LoRA-SFT and GRPO updates are approximately orthogonal. Thus, parameter addition can preserve both contributions because the two updates do not compete for the same parameter directions.

Representation-level view. A logit-lens probe (nostalgebraist, 2020) places the language decision in the last third of the network: for Chinese, Japanese, and Thai, English-only SFT first crosses the target-script threshold at layers 18/19/22 of 28 and M-GOAL crosses at the same layers, whereas GRPO pushes target-script mass below the base model (Appendix J). Language consistency thus behaves like a late output-level decision rather than a rewiring of the intermediate computation.

Direct addition suffices. Across 26 configurations of TIES (Yadav et al., 2023), DARE (Yu et al., 2024), model soup (Wortsman et al., 2022), SLERP (Goddard et al., 2024), and weighted task arithmetic, the best tuned alternative (DARE-TIES, 57.5/76.7) does not reliably improve on untuned direct addition (57.8/75.2), while soup and SLERP fall far behind (Appendix I).

5.4 Controlling the accuracy–consistency trade-off

Equation 4 is the case $\alpha = 1$ of $\theta = \theta_{\mathrm{GRPO}} + \alpha\,\Delta_{\mathrm{LoRA\text{-}SFT}}$. Sweeping α (Table 9) shows the two capabilities are not locked to a single exchange rate: accuracy is fully preserved up to $\alpha = 0.5$, consistency rises steeply between 0.75 and 1.0, and beyond 1.0 it saturates while only accuracy is lost. The coefficient can thus be chosen on held-out data for a deployment's preference; we report $\alpha = 1$ throughout, fixed once rather than tuned per language or benchmark. The residual cost at $\alpha = 1$ also delimits our orthogonality claim: $\Delta_{\mathrm{LoRA\text{-}SFT}}$ is not a pure consistency direction but also carries a weaker task-solving component that dilutes the stronger GRPO solution (Appendix H). We claim approximate functional orthogonality between the two behaviors, not purity of either update.

α	Acc	LC
0.00 (GRPO only)	62.9	12.7
0.25	62.9	12.7
0.50	63.2	15.1
0.75	61.6	41.5
1.00 (M-GOAL)	57.8	75.2
1.25	55.3	83.6
1.50	54.1	83.8

Table 9: Merge-coefficient sweep on MGSM (Qwen3-1.7B-Base), eight-language averages, for $\theta = \theta_{\mathrm{GRPO}} + \alpha\,\Delta_{\mathrm{LoRA\text{-}SFT}}$.

6 Conclusion

We showed that multilingual reasoning post-training can be improved using English-only data. Under English-only training, SFT primarily improves language consistency, whereas GRPO improves answer accuracy; sequential training in either order overwrites part of the previously acquired capability. We find that these two capabilities correspond to approximately orthogonal update directions, which allows them to be composed through parameter addition. M-GOAL implements this idea by merging an English LoRA-SFT adapter into an English GRPO checkpoint trained from the same base model, raising MGSM accuracy from 39.1 to 57.8 and language consistency from 17.1 to 75.2 without translated reasoning supervision, with the merge coefficient exposing a controllable frontier between the two. Multilingual reasoning systems can therefore obtain both stronger accuracy and target-language reasoning without constructing language-specific reasoning corpora.

Limitations

Our study has several limitations. First, although we evaluate across multiple benchmarks

and model families, the experiments are still limited to a small set of open-weight base models and moderate model scales. It remains unclear whether the same update-space orthogonality holds for much larger models, instruction models, or proprietary reasoning systems.

Second, our language coverage is limited. Our MGSM evaluation covers eight languages, and our translated-GRPO analysis focuses on Spanish, Chinese, and Thai. These languages do not fully represent the diversity of low-resource languages, morphologically rich languages, or languages with substantially different scripts and cultural contexts.

Finally, language consistency is measured with automatic language identification. While this metric captures whether the generated reasoning is primarily in the input language, it does not measure fluency, cultural appropriateness, factual faithfulness, or whether the reasoning is semantically equivalent across languages.

Ethical Considerations

This work aims to improve multilingual reasoning using English-only post-training data. All models, datasets, and benchmarks used in this work are publicly available research artifacts. We use them in accordance with their respective licenses or terms of use, only for research purposes, and do not redistribute the original artifacts. A potential positive impact is improved accessibility: non-English users may receive reasoning traces in their own language without requiring costly translated reasoning supervision for every target language. The method may also reduce the cost and noise introduced by large-scale translation pipelines.

However, stronger multilingual reasoning can also amplify general risks of language models across more languages. These risks include generating fluent but incorrect explanations, increasing user over-trust, or assisting harmful tasks in non-English languages. Language consistency should not be interpreted as a guarantee of factual correctness, safety, or cultural appropriateness. In addition, because the post-training data are English-only, the model may still inherit English-centric assumptions even when it reasons in a target language.

Our experiments use public benchmark datasets and machine-translated prompts for analysis. We do not use private user data or collect human-subject data. Before deployment in high-stakes settings, multilingual reasoning models should undergo language-specific evaluation for factuality, safety, bias, and cultural appropriateness.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New dataset and models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065. Association for Computational Linguistics.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, et al. 2025. SmolLM3: smol, multilingual, long-context reasoner. `https://huggingface.co/blog/smollm3`. Accessed 2026-05-24.

Josh Barua, Seun Eisape, Kayo Yin, and Alane Suhr. 2026. Long chain-of-thought reasoning across languages. In *International Conference on Learning Representations*.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xinnian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and Zhoujun Li. 2025. XCOT: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23550–23558.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. 2025. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

DeepSeek-AI. 2026. Deepseek-v4-flash. https://api-docs.deepseek.com/. Accessed: 2026-05-24.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, et al. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's MergeKit: A toolkit for merging large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485, Miami, Florida, US. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Hugging Face. 2025. Open r1: A fully open reproduction of DeepSeek-R1. https://github.com/huggingface/open-r1.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *International Conference on Learning Representations*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Jiwoo Kim, Youngbin Kim, Ilwoong Baek, JinYeong Bak, and Jongwuk Lee. 2023. It ain't over: A multi-aspect diverse math word problem dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14984–15011, Singapore. Association for Computational Linguistics.

Qiming Li, Xiaocheng Feng, Yixuan Ma, Zekai Ye, Ruihan Chen, Xiachong Feng, and Bing Qin. 2025. Unlocking multilingual reasoning capability of llms and lvlms through representation engineering. *arXiv preprint arXiv:2511.23231*.

Junxiao Liu, Zhijun Wang, Yixiao Li, Zhejian Lai, Liqian Huang, Xin Huang, Xue Han, Junlan Feng, and Shujian Huang. 2026. Self-improving multilingual long reasoning via translation-reasoning integrated training. *arXiv preprint arXiv:2602.05940*.

nostalgebraist. 2020. Interpreting GPT: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. LessWrong blog post.

OpenAI. 2024. Multilingual massive multitask language understanding (MMMLU). https://huggingface.co/datasets/openai/MMMLU. Professional human translations of the MMLU test set into 14 languages.

Cheonbok Park, Jeonghoon Kim, Joosung Lee, Sanghwan Bae, Jaegul Choo, and Kang Min Yoo. 2025. Cross-lingual collapse: How language-centric foundation models shape reasoning in large language models. *arXiv preprint arXiv:2506.05850*.

Jinghui Qin, Xiaodan Liang, Yining Hong, Jianheng Tang, and Liang Lin. 2021. Neural-symbolic solver for math word problems with auxiliary tasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5870–5881. Association for Computational Linguistics.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10015–10027, Bangkok, Thailand. Association for Computational Linguistics. ArXiv:2401.06838.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36.

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Fei Huang, Junyang Lin, Fei

Huang, and Jingren Zhou. 2025. PolyMath: Evaluating mathematical reasoning in multilingual contexts. *arXiv preprint arXiv:2504.18428*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, pages 7093–7115.

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*.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57755–57775. PMLR.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Kaiyu Huang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2026. Think natively: Unlocking multilingual reasoning with consistency-enhanced reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20766–20783, San Diego, California, United States. Association for Computational Linguistics.

Raoyuan Zhao, Yihong Liu, Hinrich Schuetze, and Michael A. Hedderich. 2026. A comprehensive evaluation of multilingual chain-of-thought reasoning: Performance, consistency, and faithfulness across languages. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5223–5247, Rabat, Morocco. Association for Computational Linguistics.

Wei Zhao, Mingyue Shang, Yang Liu, Liang Wang, and Jingming Liu. 2020. Ape210K: A large-scale and template-rich dataset of math word problems. *arXiv preprint arXiv:2009.11506*.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8411–8423, Bangkok, Thailand. Association for Computational Linguistics.

A Translation Prompts

A.1 Math Problems

We use the following prompt to translate English math word problems into the target language.

```
You are a literal translator for math word problem
↪  statements.
Translate the following English problem to {lang_name}.
RULES:
- Faithful and literal. Do NOT add reasoning, explanation,
  or commentary of your own.
- Preserve EXACTLY: numbers, math symbols (+, -, *, /, =,
↪  %, ^),
  LaTeX (\frac, \sqrt, \boxed, etc.), currency symbols(
      $, ¥, €, £), unit abbreviations (km, kg, hr, ...),
  single-letter variables (x, y, n, ...), people's names
  (kept in Latin), well-known place names (kept in Latin).
- Translate every English narrative word to {lang_name}.
  Do not leave common nouns/verbs/adverbs in English just
  because they are capitalized at sentence start.
- Output ONLY the translated problem statement -- no code
↪  fences,
  no prefix like "Here is the translation:".
```

A.2 Math Solutions

We use the following prompt to translate English math word solutions into the target language.

```
You are a literal translator for math word problem
↪  solutions. Translate the following English solution to
↪  {lang_name}.
RULES:
- Faithful and literal. Do NOT add reasoning of your own;
↪  do NOT shorten; translate every sentence.
- Preserve EXACTLY: numbers, math symbols, all LaTeX
↪  commands and their contents, especially \boxed{...} -
↪  the content inside the FINAL \boxed{} is the
↪  ground-truth answer and MUST remain identical to the
↪  source. Currency symbols, units, single-letter math
↪  variables, people's names (Latin), place names (Latin)
↪  - all preserved.
- Translate every English narrative word to {lang_name}.
- Output ONLY the translated solution text. Do not output
↪  anything before/after it.
```

A.3 ECQA Multiple-Choice Questions

For ECQA, we apply the translation prompt after the four-option resampling described in Appendix D. Each translated instance therefore contains one question and four answer options indexed as A–D.

```
You are a literal translator for multiple-choice question
↪  prompts.
Translate the following English question (with its four
↪  answer options)
to {lang_name}.
RULES:
- Faithful and literal. Do NOT add reasoning, explanation,
  or commentary of your own.
- Preserve EXACTLY: the letter labels "A.", "B.", "C.", "D."
  at the start of each option line. Do not translate the
  ↪  letters.
- Preserve EXACTLY: all numbers, percentages, people's
↪  names
  (kept in Latin), well-known place names (kept in Latin).
- Translate the question text and the text of every option
↪  into
  natural {lang_name}.
- Keep the same line structure: one question line, then
↪  four option
  lines (one option per line).
- Output ONLY the translated question + four options -- no
↪  code
  fences, no prefix like "Here is the translation:", no
  ↪  closing remark.
```

B Qualitative Example of Capability Composition

Model	Acc	LC	P_{zh}	Pred.
Base	0	0	0.001	None
LoRA-SFT	0	1	1.000	-29
GRPO	1	0	0.002	125
M-GOAL	1	1	0.992	125

Table 10: Qualitative MGSM example in Chinese. Base fails both objectives. LoRA-SFT restores language consistency only; GRPO restores accuracy only; merging the two obtains both.

Table 10 shows a Chinese MGSM example where the two capabilities separate cleanly. LoRA-SFT fixes language consistency but not accuracy, GRPO fixes accuracy but not language consistency, and their merge obtains both.

Question.

```
一个商人想要在两种购买计划中做出购买选择：价值 5,000
↪  美元的珠宝和价值
8,000 美元的电子产品。他的财务顾问推测，珠宝市场将上涨
↪  2.5\%，而电子产品
市场在同一个月内也将上涨 1.2\%。
↪  如果这个商人期望通过做出选择，在本月末
获得最大的利润，则利润会是多少？\\
Gold Answer: $125$.
```

Base output.

```
1. The profit from the jewelry purchase would be 5000 *
↪  2.5% = $125.
2. The profit from the electronic gadget purchase would be
↪  8000 * 1.2% = $96.
3. The total profit would be $125 + $96 = $221.
4. So, the profit from the purchase of the electronics
↪  would be greater
   than the profit from the purchase of the jewelry.
\boxed{221}
```

LoRA-SFT output.

```
珠宝和电子产品分别上涨的百分比是 2.5\% 和 1.2\%，
↪  所以珠宝的预期价格上涨
2.5\% $\times$ 5,000 美元 = 125 美元，
↪  而电子产品的预期价格上涨
1.2\% $\times$ 8,000 美元 = 96 美元。因此，
↪  通过选择购买电子产品，这个商人可以
期待获得更多的利润，即 96 美元 $-$ 125 美元 = $-29$
↪  美元的损失。
\boxed{-29}
```

GRPO output.

```
1. Calculate the new value of the jewelry after the
↪  increase:
   5000 * 1.025 = 5125 dollars.
2. Calculate the new value of the electronics after the
↪  increase:
   8000 * 1.012 = 8096 dollars.
3. Determine the profit from the jewelry:
   5125 - 5000 = 125 dollars.
4. Determine the profit from the electronics:
   8096 - 8000 = 96 dollars.
```

```
5. Compare the profits:
   125 dollars from jewelry vs. 96 dollars from electronics.
Therefore, the final answer is: \boxed{125}
```

M-GOAL output.

```
珠宝和电子产品的当前价格分别为 5,000 美元和 8,000 美元。
↪ 珠宝市场预期上涨
2.5\%，电子产品市场预期上涨 1.2\%。
↪ 我们需要计算每种产品的预期价格增长，
并比较最终利润。
对于珠宝：当前价格为 \$5,000；预期价格上涨为 2.5\%；
↪ 预计价格为
\$5,000 $\times$ 1.025 = \$5,125。
对于电子产品：当前价格为 \$8,000；预期价格上涨为 1.2\%；
↪ 预计价格为
\$8,000 $\times$ 1.012 = \$8,096。
珠宝的预期利润为 \$5,125 $-$ \$5,000 = \$125。
↪ 电子产品的预期利润为
\$8,096 $-$ \$8,000 = \$96。比较利润，珠宝的预期利润更高。
↪ 因此，为了在本月末
获得最大的利润，商人应该选择购买珠宝。
最终利润为 $\boxed{125}$。
```

C Additional Translated-GRPO Experiments

	10% Trans	50% Trans	100% Trans
Spanish Acc	67.1	63.9	64.1
Chinese Acc	66.0	64.7	61.1
Thai Acc	65.2	68.1	67.1
Spanish LC	90.4	91.2	95.0
Chinese LC	75.5	84.6	27.0
Thai LC	86.5	89.2	91.0

Table 11: Effect of translated GRPO data on MGSM for Qwen3-1.7B. Adding only 10% non-English prompts sharply increases language consistency across the three target languages while preserving accuracy.

We compare mixtures with 10%, 50%, and 100% translated (non-English) prompts across three target languages: Spanish, Chinese, and Thai (Table 11). These languages cover a Latin-script non-English language, a high-resource non-Latin language, and a comparatively lower-resource non-Latin language, respectively. A small translated fraction is already highly effective. With only 10% non-English prompts, LC rises to 90.4 for Spanish, 75.5 for Chinese, and 86.5 for Thai, while accuracy is preserved (e.g., 67.1 on Spanish). The 100% translated setting is not uniformly better: Chinese LC collapses, while Spanish and Thai remain high. This indicates a language-specific instability rather than a general failure of translated GRPO. We hypothesize that monolingual translated GRPO may reinforce a single-language response mode instead of conditional language consistency, especially when the target language has strong pre-existing model priors. In Chinese, continued GRPO may first move the model away from an English-default mode toward a transient language-consistent regime, but later push it into a Chinese-default mode, hurting multilingual evaluation.

C.1 Translation Cost, Quality Control, and Task Dependence

Constructing the translated-GRPO conditions required translating 19,759 source problems separately into Spanish, Chinese, and Thai with DeepSeek-V4-Flash (DeepSeek-AI, 2026). We verified the language of every translated prompt with fastText and manually inspected a random subset. Quality was not uniform across languages: roughly 30% of the Thai items failed our automatic checks and required retranslation, best-of-n resampling with an NLLB fallback, or manual repair.

The translated fraction needed to restore language consistency is also task dependent. On mathematical reasoning a 10% translated mixture is already close to saturation (Table 11). On the ECQA→MMMLU setting the same 10% mixture raises Spanish LC only from 13.4 to 14.3, and the mixture must approach 50% before LC comes near M-GOAL's 74.6 on the same split. A practitioner adopting translated GRPO therefore has to search the mixture ratio per task and per language, in addition to constructing and auditing the translations.

For completeness we also report the head-to-head comparison on the subset of languages for which we trained translated-GRPO models. Averaged over Spanish, Chinese, and Thai, translated GRPO reaches 66.1 Acc / 84.1 LC and M-GOAL reaches 64.5 / 84.3: close on both axes, with translated GRPO slightly ahead on accuracy and M-GOAL slightly ahead on language consistency. The two settings are not matched in supervision, since translated GRPO uses target-language prompts in all three languages while M-GOAL uses none.

D Benchmark and Dataset Details

We summarize the benchmarks and datasets used in our experiments. These include multilingual math benchmarks, an English commonsense-reasoning dataset for post-training, and a multilingual non-mathematical benchmark for evaluating cross-domain transfer. As summarized in Table 12, our evaluation covers both high-resource and comparatively lower-resource languages across multi-

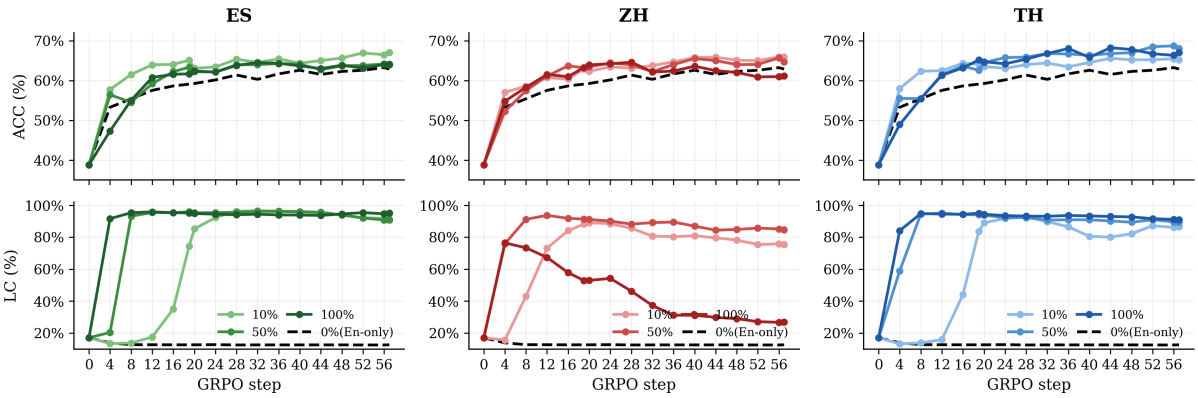


Figure 10: GRPO training dynamics under different non-English data fractions (10%/50%/100%) across target languages. Top row: accuracy; bottom row: language consistency. A small (10%) translated fraction sharply raises language consistency while largely preserving accuracy, whereas fully non-English training is not uniformly beneficial.

ple language families and writing systems, including Germanic, Romance, Sinitic, Japonic, Slavic, Semitic, Tai-Kadai, Bantu, and Austronesian languages.

MGSM. MGSM is a multilingual grade-school math benchmark consisting of parallel translations of GSM8K-style problems across multiple languages (Shi et al., 2023). We use MGSM as our primary evaluation benchmark because it directly tests both answer accuracy and language consistency in multilingual mathematical reasoning. Since each problem is aligned across languages, MGSM also allows us to compare how English-only post-training affects the same reasoning problem under different input languages. We evaluate on 247 examples per language after removing three examples that overlap with GSM8K few-shot demonstrations.

MSVAMP. MSVAMP is a multilingual arithmetic word-problem benchmark derived from SVAMP and used to evaluate cross-lingual mathematical reasoning (Chen et al., 2024). Compared with MGSM, MSVAMP provides a different distribution of short arithmetic problems, making it useful for testing whether the accuracy–language-consistency trade-off is specific to one multilingual math dataset or persists across word-problem formulations.

PolyMath-low. PolyMath-low is a multilingual mathematical reasoning benchmark designed to cover a broader set of languages and mathematical problem types (Wang et al., 2025). We use its low-resource-language split to test whether the observed effects extend beyond the relatively high-resource languages covered by MGSM. This benchmark is therefore useful for evaluating whether M-GOAL improves both reasoning accuracy and language consistency under more diverse multilingual conditions.

ECQA. ECQA is an explainable commonsense question-answering dataset that pairs multiple-choice commonsense questions with natural-language explanations (Aggarwal et al., 2021). We use ECQA as the English post-training source for the non-mathematical ablation. To make the ECQA format consistent with our four-option prompting and translation template, we resample each original example by randomly removing one incorrect answer option together with its corresponding explanation. The correct option is always retained. The remaining four options are then re-indexed as A–D, and the associated explanations are re-aligned to the new option labels. This preprocessing preserves the original question and gold answer while converting ECQA into a four-choice format used throughout our ECQA post-training experiments. This allows us to test whether the separation between language-consistency gains from SFT and accuracy gains from GRPO is specific to mathematical reasoning, or whether it also appears in a commonsense reasoning setting.

MMMLU. MMMLU is a multilingual extension of MMLU that evaluates broad knowledge and reasoning ability across multiple subjects and languages (OpenAI, 2024). We use MMMLU as the cross-lingual evaluation benchmark after English-

Language	MMMLU	MGSM	MSVAMP	PolyMath-low
English	✓	✓	✓	✓
Chinese	✓	✓	✓	✓
German	✓	✓	✓	✓
Spanish	✓	✓	✓	✓
French	✓	✓	✓	✓
Japanese	✓	✓	✓	✓
Arabic	✓			
Indonesian	✓			
Russian			✓	✓
Thai		✓		
Swahili	✓	✓	✓	✓

Table 12: Language coverage across evaluation benchmarks. Checkmarks indicate that the language is included in the corresponding benchmark.

only ECQA post-training. Because MMMLU covers non-mathematical domains, it provides a stress test for whether M-GOAL's accuracy–consistency improvements transfer beyond math word problems.

C-MGSM. MGSM and the other multilingual math benchmarks we use are built by translating English source problems, so their non-English items may carry translationese. C-MGSM inverts this direction: it is built from natively authored Chinese elementary-school word problems, which are then translated into the remaining evaluation languages, so that Chinese rather than English is the source language. We report seven-language averages.

CM17K, Ape210K, and DMath. To evaluate on problems that are never translated at all, we additionally use two natively Chinese math word-problem datasets and one Korean split. CM17K (Qin et al., 2021) contains more than 17K Chinese problems spanning arithmetic, one-unknown linear, one-unknown non-linear, and equation-set types. Ape210K (Zhao et al., 2020) is a large-scale, template-rich collection of Chinese math word problems. DMath (Kim et al., 2023) is a multi-aspect math word-problem dataset released in parallel English and Korean; we evaluate on its Korean split. Because each is monolingual, these three isolate target-language performance without any cross-lingual parallelism, and their Acc/LC in Table 4 is computed exactly as for MGSM.

E Behavioral Independence Details

Figure 11 gives an alternative view of the behavioral-independence test. Each point compares the observed overlap between the LoRA-SFT language-consistency improvement set and the GRPO accuracy improvement set against the overlap expected under independence. Points close to the diagonal indicate nearly independent improvement sets.

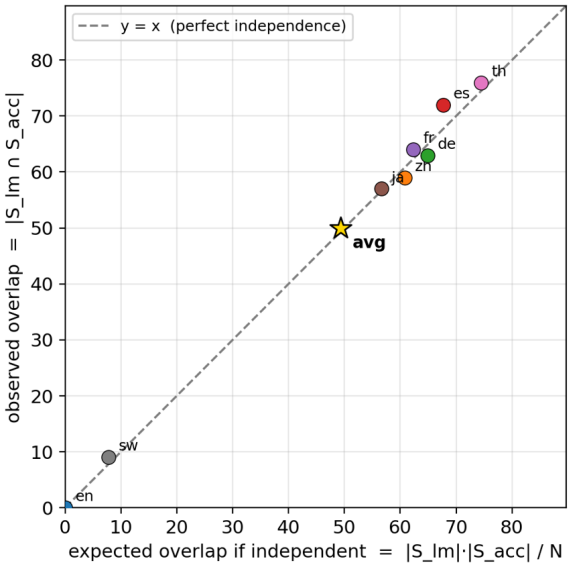


Figure 11: Behavioral independence between LoRA-SFT language-consistency gains and GRPO accuracy gains. Each point shows one language, comparing the observed joint-improvement count against the count expected under independence, $|S_{\mathrm{LC}}||S_{\mathrm{Acc}}|/N$. Points close to the diagonal indicate near-independent improvement sets.

Table 13 reports the exact counts. Here, S_{LC} is the set of examples whose language consistency is improved by LoRA-SFT over the base model, and S_{Acc} is the set of examples whose answer accuracy is improved by GRPO over the base model. For English, $|S_{\mathrm{LC}}| = 0$ because the base model already has perfect English language consistency, so the ratio is undefined.

F Additional Experiment Details

F.1 OpenR1 Dataset Filtering

We construct the OpenR1 subset by applying four automatic filters to the default split, as summarized in Table 14.

These filters reduce the dataset to examples suitable for text-only, reward-based training: the retained problems have manageable reasoning traces, deterministic numerical targets, and no reliance on missing visual context. The final stage further selects problems with nontrivial but learnable difficulty, leaving a smaller subset with stronger training signal.

Lang.	N	$\lvert S_{\mathrm{LC}}\rvert$	$\lvert S_{\mathrm{Acc}}\rvert$	Obs.	Exp.
en	247	0	93	0	0.00
zh	247	203	74	59	60.82
de	247	193	83	63	64.85
es	247	209	80	72	67.69
fr	247	208	74	64	62.32
ja	247	212	66	57	56.65
th	247	239	77	76	74.51
sw	247	127	15	9	7.71
Avg.	247	173.9	70.2	50	49.32
Overall	1976	1391	562	400	395.62

Table 13: Per-language counts for the behavioral-independence test. The observed joint-improvement counts closely match the independence expectation. The "Avg." row averages the per-language values, while the "Overall" row pools all languages before applying the $|S_{LC}| \cdot |S_{Acc}|/N$ formula; the two Exp. values therefore use different aggregation and are not related by a factor of 8.

Stage	Filter	Remain.
0	Default OpenR1 split	∼94K / 100%
1	Keep central reference-trace lengths under the Qwen3-1.7B tokenizer (P10–P90; ∼1.8K–11.7K tokens)	∼75K / ∼80%
2	Require numerically parseable gold answers, either floats or clean fractions	∼35K / ∼37%
3	Remove geometry problems, which may require unavailable image context	∼28K / ∼30%
4	Keep medium-difficulty examples under eight Qwen3-1.7B-Base rollouts: $1 \le c \le 6$ correct	∼12K / ∼13%

Table 14: OpenR1 filtering pipeline. Ratios are approximate and computed relative to the original split. The final filter discards unsolved examples ($c = 0$) and nearly saturated examples ($c \ge 7$), where c is the number of correct responses among eight sampled rollouts.

F.2 Full Training Objectives

For completeness, we include the standard SFT and GRPO objectives used in our experiments.

SFT objective. In SFT, the model is trained on question–reasoning–answer demonstrations. Given a question q and a target reasoning–answer sequence $o = (o_1, \ldots, o_{|o|})$, SFT minimizes token-level negative log-likelihood:

$$\mathcal{L} = -\mathbb{E}_{(q,o)\sim\mathcal{D}_{\mathrm{SFT}}}\left[\frac{1}{|o|}\sum_{t=1}^{|o|}\log \pi_\theta(o_t \mid q, o_{<t})\right]. \tag{5}$$

GRPO objective. For each question q, GRPO samples a group of N responses $\{o_i\}_{i=1}^N$ from the old policy $\pi_{\theta_{\mathrm{old}}}$. The token-level probability ratio is

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(o_{i,t} \mid q, o_{i,<t})}. \tag{6}$$

GRPO maximizes the clipped policy objective

$$\mathcal{J} = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(g_{i,t}(\theta) - \beta D_{i,t}^{\mathrm{KL}}\right)\right], \tag{7}$$

where

$$g_{i,t}(\theta) = \min\left(\rho_{i,t}(\theta)A_i,\ \bar{\rho}_{i,t}(\theta)A_i\right). \tag{8}$$

$$\bar{\rho}_{i,t}(\theta) = \mathrm{clip}(\rho_{i,t}(\theta), 1-\epsilon, 1+\epsilon).$$

The advantage is computed by normalizing rewards within the sampled group:

$$A_i = \frac{r_i - \bar{r}}{s_r}, \qquad r_i = R(q, o_i), \tag{9}$$

where $\bar{r}$ and s_r are the mean and standard deviation of $\{r_j\}_{j=1}^N$.

F.3 Training Setup

Table 15 reports the hyperparameters used for GRPO, full SFT, and LoRA-SFT. These are the configurations used in our main experiments.

Hyperparameter	Value
GRPO	
Learning rate	1×10^{-5}
KL loss / coefficient	True / 1×10^{-4}
Rollouts per prompt	8
Temperature	1.0
Epochs	3
Full SFT	
Learning rate	1×10^{-4}
Epochs	3
LoRA-SFT	
Rank r	64
Scaling α	128
α/r	2
Learning rate	1×10^{-4}
Epochs	3

Table 15: Main post-training hyperparameters.

G Accuracy-Only GRPO Reward Control

To confirm that the language-consistency collapse under English-only GRPO is not induced by the LC term of Equation 3, we retrained GRPO with $R(q, o) = \mathrm{Acc}(q, o)$, keeping the training data,

GRPO reward	Acc	LC
Acc only	63.2	12.6
$0.5\,\mathrm{Acc} + 0.5\,\mathrm{LC}$ (Eq. 3)	62.9	12.7

Table 16: Accuracy-only GRPO control on MGSM (Qwen3-1.7B-Base), trained with identical data and hyperparameters. Removing the language-consistency term changes neither objective.

the number of rollouts, the learning rate, the KL coefficient, and the number of epochs identical. The final MGSM averages are 63.2 Acc / 12.6 LC against 62.9 / 12.7 for the mixed reward (Table 16); non-English LC is near zero for both runs at every evaluated checkpoint.

Why the LC term is inert is visible in the training statistics rather than in the final scores. With English-only prompts, the rollouts within a group are almost always English, so $\mathrm{LC}(q, o_i)$ takes the same value for nearly every member of the group. Because GRPO normalizes rewards within the group, a component that is constant across the group contributes nothing to the advantage A_i and therefore does not shape the update. The LC reward becomes an effective signal only when the prompt distribution makes the target language vary, which is what the translated-GRPO conditions of Section 5.2 supply.

H Merge-Coefficient Sweep

Table 9 reports eight-language MGSM averages for $\theta = \theta_{\mathrm{GRPO}} + \alpha\,\Delta_{\mathrm{LoRA\text{-}SFT}}$ with $\alpha \in \{0, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5\}$. The curve has three regimes. For $\alpha \leq 0.5$ the merged model is behaviorally close to GRPO: accuracy is unchanged, and marginally higher at $\alpha = 0.5$, while LC has barely moved. Between $\alpha = 0.75$ and $\alpha = 1.0$, LC rises from 41.5 to 75.2 while accuracy falls by 3.8 points, which is where the adapter begins to dominate the output-language decision. Beyond $\alpha = 1.0$, LC saturates near 84 while accuracy continues to decay, so larger coefficients are not useful. We use $\alpha = 1$ for all M-GOAL results in the paper, fixed once rather than tuned per language or per benchmark, so that the reported numbers are not the product of a per-setting search.

The residual accuracy cost at $\alpha = 1$ delimits what our orthogonality claim asserts. $\Delta_{\mathrm{LoRA\text{-}SFT}}$ is not a pure language-consistency direction: alongside the dominant consistency component it carries a smaller task-solving and response-style component learned from the SFT demonstrations. That component is weaker than the GRPO solution, so adding it perturbs GRPO's task-solving behavior. Consistent with this dilution account, across the eight MGSM languages the per-language accuracy drop tracks the GRPO–SFT competence gap rather than the size of the consistency gain. Our claim is therefore approximate functional orthogonality between the consistency behavior and the task-solving behavior, not that either update is a pure single-capability vector.

I Comparison with Other Model-Merging Methods

Direct parameter addition is the simplest way to compose θ_{GRPO} and $\Delta_{\mathrm{LoRA\text{-}SFT}}$, but the model merging literature offers alternatives. We evaluated 26 configurations in total, covering TIES (Yadav et al., 2023) at several trim densities, DARE-linear and DARE-TIES (Yu et al., 2024) at several drop rates, uniform model soup (Wortsman et al., 2022), SLERP as implemented in MergeKit (Goddard et al., 2024), and weighted task arithmetic with several weightings.

Merging rule	Acc	LC
Direct addition (M-GOAL, untuned)	57.8	75.2
DARE-TIES (best of 25 tuned)	57.5	76.7
Model soup	50.3	32.3
SLERP	50.3	31.9

Table 17: Representative results from the merging-rule comparison on MGSM (Qwen3-1.7B-Base, eight-language averages). Direct addition requires no tuning and is not reliably improved upon by the best tuned alternative.

Table 17 summarizes the outcome. The best tuned alternative, DARE-TIES, reaches 57.5 Acc / 76.7 LC, which we do not regard as a reliable improvement over untuned direct addition (57.8 / 75.2) given that we report single runs. Interpolation-based rules that average the two endpoints instead of adding an update, namely model soup and SLERP, are clearly worse on both axes, which is expected because they pull the model halfway back toward the SFT solution rather than composing the two updates.

All merges were computed in fp32. This matters in practice: the GRPO update $\Delta_{\mathrm{GRPO}} = \theta_{\mathrm{GRPO}} - \theta_{\mathrm{base}}$ has a small per-parameter magnitude, and carrying out the arithmetic in bf16

rounds part of it away, which changes the measured behavior of every merging rule and makes comparisons across rules unreliable.

J Logit-Lens Analysis of Where the Language Decision Emerges

To obtain representation-level evidence for why an English-only SFT adapter can restore non-English reasoning, we apply a logit-lens probe (nostalgebraist, 2020) to the first 20 generated tokens of each model's own MGSM output. For every transformer layer we pass the residual stream through the final RMSNorm and the language-model head, and record the probability mass assigned to tokens in the target script.

For Chinese, Japanese, and Thai the base model never crosses our target-script threshold at any layer, ending at 0.114, 0.048, and 0.059 respectively. After English-only SFT the target-script mass crosses the threshold at layers 18, 19, and 22 of 28, reaching 0.772, 0.881, and 0.978. M-GOAL crosses at the same layers and reaches 0.894, 0.744, and 0.895. English-only GRPO moves in the opposite direction, ending far below the base model, for example 0.0045 versus 0.114 in Chinese.

Two observations follow. First, the language decision is made late: the middle layers look similar across all four models, and target-language mass appears only in the final third of the network. This is consistent with language consistency being carried by a small update that need not reorganize the intermediate computation, and therefore with its being separable into a LoRA adapter. Second, M-GOAL reproduces the SFT model's layerwise profile while retaining GRPO's answers, which is what the composition hypothesis predicts.

Because each model is probed on the tokens it generates itself rather than on a shared forced-decoding prefix, these trajectories are not a controlled comparison of hidden representations, and we present them as descriptive support rather than as a causal mechanism.