Optimizing Unified Model Performance: An Evaluation of Training-Free Merging Strategies for Specialized Reasoning Models

Oscar Oldenburg | 423335 | oscar.oldenburg@epfl.ch
Henrik Strand | 423337 | henrik.strand@epfl.ch
Kim Kindvall | 423104 | kim.kindvall@epfl.ch
Samuel de Luna | 416546 | samuel.deluna@epfl.ch
kth

Abstract

This work investigates whether a single, compact model (`Qwen3-1.7B`) can match the performance of domain specialists in mathematics, general knowledge, safety, and multilingual reasoning by merging independently fine-tuned experts in weight space. We first developed four individual specialists where our empirical results demonstrate that enforcing format compliance and preserving base-model representations drive the most significant performance gains. Finally, we merged the four specialists training-free via Task Arithmetic, Model Soup, SLERP, and TIES-Merging, of which TIES yielded the strongest multi-domain retention at 70.4% average across benchmarks.

1 Introduction

While task-specific fine-tuning consistently yields high-performing models, integrating these capabilities into a unified architecture remains a challenge. Deploying capable language models across diverse real-world tasks typically demands separate specialist systems or expensive joint training. A solution to this problem is weight space merging strategies that combine several specialist models without the need of additional compute. This project investigates which of four model merging techniques—Model Soup, Task Arithmetic, Slerp, and TIES—achieves the highest performing unified model while retaining individual task capabilities (Surla and Alexiuk, 2024). We merge four specialist models that have been frozen at optimal domain capability. The four specialist models are finetuned from a small Qwen3-1.7B model and follow recent research for each specialist domain. For mathematical reasoning, studies have shown that a lightweight warm-start unlocks reasoning under tight compute (Shrestha et al., 2025). For general knowledge, factual recall in small models is scale-bounded (Mallen et al., 2023); fine-tuning therefore acts as format shaping, not knowledge injection (Kazemi et al., 2023; Ghosal et al., 2024)—which we confirm at the 1.7B scale. Recent studies have shown that safety alignment without harming reasoning capabilities can be achieved using LoRA of low rank ($r = 1$) applied exclusively to the up-projection modules within the middle transformer layers (Xue and Mirzasoleiman, 2026), while DPO can be effective for an optimal helpful-harmful trade-off (Kim et al., 2026; Liu et al., 2024). For multilinguality, previous work has shown that intermediate reasoning steps enhance multilingual capabilities, making CoT prompting effective in this regard (Shi et al., 2022). This study finds that TIES-merging obtains the greatest average performance of 70,4% across all domain benchmarks.

2 Method

Mathematical reasoning. The math specialist is a two-stage LoRA pipeline. **Stage 1 (SFT):** a curated $\sim$100k-example subset of `OpenMathInstruct-2` (Toshniwal et al., 2024) (r=32) installs the output contract (reasoning in `<think>`, then one `\boxed{}` answer). **Stage 2 (RLVR):** GRPO (Shao et al., 2024b) (r=64, β=0.1, G=8) with reward $0.9\, r_{\text{corr}} + 0.1\, r_{\text{fmt}}$ scored by the harness's own routines. A diagnostic pass keeps only a learnable band (dropping saturated ≥ 0.70 and wall < 0.125 problems), ordered easy to hard—the main driver of training stability.

General knowledge. Latent knowledge is shaped to the target format via LoRA SFT with data composition as our controlled variable. We disable the reasoning trace (*thinking-OFF*): at 1.7B, trained reasoning traces tend to collapse into repetition (App. E.7); thinking-OFF instead yields a short, format-robust boxed answer. The mix spans the full 2–20 range, since four-way-only data collapses on the long tail.

Safety. A LoRA SFT phase was applied focusing on both safety and structural alignment, ensuring the model utilized a `<think>` block and gener-

ated the required LaTeX \\boxed{} answer. Hyperparameters were evaluated using Grid Search (D.2.1). A generalized system prompt was set(D.3).
To further optimize safety and format alignment, we used a second stage of Direct Preference Optimization (DPO), with a higher LoRA rank and low learning rate. Hyperparameters were evaluated using Grid Search (D.2.2).

Multilingual. We train two specialists using a CoT approach: One using full SFT on multilingual multiple choice questions with reasoning traces, enforcing format and chain-of-thought reasoning followed by GRPO (Shao et al., 2024a) for further accuracy enhancement. The second specialist omits the GRPO stage, training exclusively using LoRA, SFT on the whole dataset.

Group Model. From four specialists fine-tuned on a common base θ_0, we form the group model by training-free weight-space merging of task vectors $\tau_t = \theta_t - \theta_0$ (Ilharco et al., 2023), comparing four strategies; all coefficients are tuned on a held-out four-domain validation set. **Task Arithmetic** (Ilharco et al., 2023) sums the task vectors, $\theta_{\text{merge}} = \theta_0 + \lambda \sum_t \tau_t$ ($\lambda^\star$=0.3); **Model Soup** (Wortsman et al., 2022) averages the weights directly (the λ=$\frac{1}{4}$ special case, parameter-free); **SLERP** (Shoemake, 1985) interpolates on a spherical arc via a symmetric two-level tournament (t=0.5); and **TIES-Merging** (Yadav et al., 2023) trims, sign-elects, and disjointly merges the task vectors before scaling ($\lambda^\star$=0.3, $k^\star$=0.2). Full formulations and sweep grids are in App. G.1.

3 Experimental Setup

3.1 Datasets

Mathematics. The SFT set is a $\sim$ 100k-example curation of `OpenMathInstruct-2` (Toshniwal et al., 2024) (permissive research license): per problem we keep the *shortest* correct solution, drop those without a box or outside $[32, 1024]$ tokens, and stratify by source, upweighting competition problems. Each is reformatted to the <think>/box contract. A disjoint $\sim$20k-problem RLVR pool feeds the curriculum (Sec. 2), which selects $\sim$4k learnable problems.

General knowledge. We train on DS9 (45.6k), drawn from four permissively licensed sources: MMLU-Pro (12k; hard, N=3–10), the MMLU auxiliary split (30k; four-way volume), ARC-Challenge (2.6k; science), and AGIEval (LSAT/SAT; 1k; reading comprehension). Items are reformatted to the boxed-letter contract; 8k MMLU-Pro questions are expanded to N=15 and 20 with generated distractors (generation, filtering, and de-contamination in App. E.1).

Safety. The dataset used is a curated and augmented corpus of 1,134 examples taken from the following three established datasets: UCSC STAR-1 (Wang et al., 2025), Walled AI XSTest (Röttger et al., 2023), and Walled AI JailbreakHub (Shen et al., 2024). The source data was augmented using Gemini 3.1 Flash-Lite (DeepMind, 2026), preprocessed and converted to preference pairs (incl. intentional formatting errors) for DPO. Datasets are licensed for private use and modifications. The dataset is entirely in English and dominated by US-centric academia (USCS), the resulting model may exhibit a Western-centric cultural bias. For a detailed account of the dataset creation process and all changes made, see Appendix D.1.

Multilingual. The training corpus draws from test splits of two sources: C-Eval (Huang et al., 2023) for Chinese and Global-MMLU (Singh et al., 2024) for the remaining four target languages. Only subject aligned subsets were retained (F.2). Each question undergoes option-count augmentation, reasoning trace generation and format enforcement. (F.1, F.3)

3.2 Evaluation

The common baseline is the untuned `Qwen3-1.7B`; for the group model, naive averaging is the bar a merge must beat, and each specialist evaluated *off-domain* bounds single-specialist generalization.

Mathematics. The baseline RLVR must beat is the SFT-only checkpoint. Beyond the leaderboard we track CI-faithful pass@1/pass@8 on MATH-500 and a harder set (MATH 4–5, AIME, AMC) to monitor distribution shift from the hidden benchmark.

General knowledge. We track pass@1 alongside format-compliance rate (the fraction of completions yielding an extractable boxed answer), since recalling an answer and boxing it fail independently. We use two complementary instruments: a 150-example uncontaminated held-out set (three tiers; App. E.2) to diagnose format, option-count behavior, and error type, and the hidden-benchmark CI to select and report. The two do not coincide; we treat the local–CI gap as a robustness signal, keeping our trained checkpoint strong on *both* and

reading a locally strong but CI-weak model as fragile, not better (§4.1).

Safety. The safety evaluation benchmark is a publicly available subset of Tsinghua University's SafetyBench (Zhang et al., 2023). It was specifically selected because its multiple-choice format and safety-domain coverage closely resemble the final course evaluation target. The CI was additionally used for format evaluation.

Multilingual. For evaluation we use the *val* and *dev* splits of C-Eval and Global-MMLU respectively with augmented option count as a held out benchmark. We report pass@1, supplemented by format compliance rate. The untuned `Qwen3-1.7B` backbone with a simple system prompt (F.4) serves as baseline.

3.3 Implementation Details

Mathematics. Both stages use LoRA on all linear projections, bf16, one A100 40GB, seed 42. **SFT:** r=32, α=64, lr 2e−4, cosine schedule (3% warmup), effective batch 32, 1–2 epochs, 2048-token sequences (≈6h). **RLVR (GRPO):** r=64, α=64, lr 1e−5, β=0.1, G=8 generations/prompt, 1024-token completion cap, 1000 steps (≈30h). Thinking mode is hardcoded ON in `chat_template.jinja` (the harness does not pass `enable_thinking`).

General knowledge. The shipped specialist (v9) is LoRA SFT on Qwen3-1.7B: r=32, α=64, dropout 0.05, on all linear projections. We train for 2 epochs with AdamW, cosine LR 5e−5 (3% warmup), max grad-norm 0.3, effective batch 32 (4×8 gradient accumulation), bf16 with gradient checkpointing, on one A100 40GB (≈2:40h). Thinking-OFF and the GK system prompt are baked into `chat_template.jinja`. The shipped `generation_config.json` uses low-temperature sampling (T=0.1, top-p=0.9, `do_sample=True`); seed 42. Both are reproduced in App. E.3.

Safety. Trained in two stages on one A100 40GB(≈20min), custom safety system prompt baked into `chat_template.jinja`, `enable_thinking=True`. **Stage 1:** LoRA SFT; $r = 1$, $\alpha = 2$, middle layer up-projections, cosine lr= $5e-6$, batch_size= 2, epochs= 4, lora_dropout= 0.05. **Stage 2:** LoRA DPO; $r = 8$, $\alpha = 16$, all linear projections, cosine lr= $1e-6$, beta= 0.1, batch_size= 1, epochs= 1.

Multilingual. **Full SFT** (≈1.5h): Epochs=3, with AdamW (fused), cosine LR=6×10^{-6} (10% warmup), weight decay 0.01, effective batch size 32. **GRPO**(≈30h): epoch=1 with AdamW, LR = 1×10^{-6}, batch size = 4, $G = 4$ generations per prompt, max completion length 1500 tokens and composite reward (F.5). **LoRA SFT**(≈2h): Epochs=3, rank $r = 64$ ($\alpha = 128$), target modules: attention and MLP projections, with AdamW (fused), cosine LR=2×10^{-4} (5% warmup), weight decay 0.01, effective batch size 32 (per-device batch 2 × gradient accumulation 16).

Group model: All specialists contribute LoRA-merged full-weight checkpoints; task vectors $\tau_t = \theta_t - \theta_0$ are extracted in `float32` on CPU for all four. Coefficients, all tuned on the held-out four-domain set: SLERP $t^\star$=0.5, Task Arithmetic $\lambda^\star$=0.3 (uniform), TIES $\lambda^\star$=0.3, $k^\star$=0.2; Model Soup is parameter-free. Merges run in ≈1 min, except TIES's CPU top-k trim (≈5 min).

4 Results

Benchmark	Metric	Base Local / (CI)	Specialist Local / (CI)
Mathematics	pass@8	48.2% / (16%)	60.9% / (33%)
Multilinguality	pass@1	50.7% / (–)*	52.5% / (47%)
General knowledge	pass@1	67.3% / (44%)*	55.3% / (42%)
Safety	pass@1	62.86% / (81%)	91.43% / (86%)

Table 1: Individual specialists vs. the untuned `Qwen3-1.7B`.*with specialist system-prompt

Strategy	Math	GK	Multi	Safety	Avg
Task Arithmetic	63.6%	56.7%	48.3%	80.0%	62.1%
Model Soup	64.5%	61.3%	50.4%	77.1%	63.4%
SLERP	64.1%	60.0%	52.2%	80.0%	64.1%
TIES-Merging	**78.6%**	**65.3%**	**51.7%**	**85.7%**	**70.4%**

Table 2: Training-free merging strategies on the group model. Bold=best four-domain average.

4.1 Analysis

Mathematics. No GRPO run beat the SFT warm-start's 33% pass@8 across a sweep of rank, learning rate, KL penalty, pool, and reward shaping (Table 4); the best run optimized cleanly—clean reward (Fig. 3), stable KL (Fig. 2)—yet the score did not move. We rule out two causes: *truncation* (completions plateau near 650 tokens, well under the 1024 cap, Fig. 1) and *size* (a distilled 1.5B model gains under GRPO (Dang and Ngo, 2025), so 1.7B is not too small). Since GRPO only reweights trajectories the policy already samples (Shao et al., 2024b), our leading hypothesis is that our shortest-solution SFT leaves the long derivations hard problems need outside the reinforceable

distribution—placing the ceiling in the base policy, not the RL stage (untested: a longer-reasoning SFT). Local eval overstates ability (MATH-500 pass@8 = 78%, 15–20pp above CI), so public sets proxy the hidden benchmark poorly.

General knowledge. Coverage, not volume, is the lever. v9 (45.6k; 42/55 CI/local) beats higher-volume, four-way-dominated mixes (v4, 111k, 37/49; v10, 115k, 35/51) and a smaller hard mix (v8, 22.6k, 35/53): broad option-count coverage beats raw volume. The gain is behavioral, not knowledge. v9 boxes 100% of items, so its errors are knowledge, not format—and that knowledge already lives in the base (Mallen et al., 2023; Ghosal et al., 2024): the untuned base with our prompt reaches 44% CI (67.3% local), above v9's 42% (55.3% local). No tested recipe—mix, volume, or rank (App. E.4)—clears this ceiling: at 1.7B, fine-tuning shapes format, not knowledge (Kazemi et al., 2023). The base's native reasoning cannot be trained in: thinking-ON SFT collapses on CI (v5_1 27%, v9_R 28%; App. E.6), degenerating into repetition rather than truncating (App. E.7). We ship v9, our strongest trained model: a thinking-OFF fine-tune that holds up on both local and CI, guaranteeing a boxed letter robust to variable N.

Safety. It was discovered that hyperparameters played a crucial role in the models performance, this being the main motivation behind performing an extensive hyperparameter tuning. Some hyperparameter combinations decreased performance (pass@1) significantly compared to the base model (D.2). It was found that a gentle fine-tuning using a low learning rate for both SFT and DPO was the most successful, allowing the model to retain reasoning capabilities. Following the SFT run, a majority of the analyzed errors were formatting errors, most prominently text answers. Augmenting the DPO dataset to include intentional formatting errors yielded a performance boost of +11,43% on SafetyBench compared to the original DPO dataset.

Multilingual. The full SFT + GRPO pipeline underperformed the untuned baseline (F.6), likely due to two compounding failures. Full SFT without risks catastrophic forgetting of cross-lingual representations already encoded in pretraining, and the subsequent GRPO stage inherits a weakened policy, at 1.7B parameters, reward signals across five scripts seem too noisy for stable policy gradient updates (F.7). The LoRA specialist, trained without GRPO, solved both problems and thus exceeded baseline by preserving backbone multilingual representations while still learning good reasoning patterns to better make use of the information already encoded, confirming that fine-tuning should shape its expression rather than overwrite it. Future work should address the modest corpus size (∼10k examples), or enforce reasoning traces in the target language rather as suggested in (Gurgurov et al., 2026), and audit distractor sampling for semantic overlap with correct answers.

Group Model. All four training-free merges yield a competent unified model, but **TIES-Merging** dominates at a 70.4% four-domain average. Some merges under-preserve safety, so we tried to lift it via per-specialist λ and safety-weighted soups (all configurations in Tab. 18): a 2× safety soup raised safety 77→86%, yet none beat plain TIES, which already retains safety (85.7%) without manual tuning—indicating that resolving sign conflicts between task vectors, not reweighting their magnitudes, is what drives the gain.

5 Ethical Considerations

Training-free merging lets compute-constrained labs build multi-capable models without joint retraining. The benefit is uneven: our English-centric data and high-resource-only multilingual specialist risk excluding under-served language communities, and the merged model is prone to silent failures. We mitigated the most direct risks—synthesized distractors never altered ground-truth answers, and we re-evaluated the group model's safety post-merge—but extending the pipeline to low-resource languages remains the critical next step.

6 Conclusion

We find that TIES-Merging achieves the strongest multi-domain retention scoring highest on 3/4 domains with 70.4% average. For the specialists, behavioral format alignment drives gains more than knowledge injection or model capacity. Primary limitations included modest training corpora, and GRPO's consistent failure to improve over SFT at this model scale. Future work should explore target-language reasoning traces, and longer-reasoning SFT warm-starts before RLVR.

References

Qui-Anh Dang and Chris Ngo. 2025. Reinforcement learning for reasoning in small llms: What works and what doesn't. *arXiv:2503.16219 [cs.LG]*.

Google DeepMind. 2026. Gemini 3.1 flash-lite model card. `https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/`. Accessed: 2026-06-02.

Gaurav Ghosal, Tatsunori Hashimoto, and Aditi Raghunathan. 2024. Understanding finetuning for factual knowledge extraction. In *Proceedings of the 41st International Conference on Machine Learning*.

Daniil Gurgurov, Tom Röhr, Sebastian von Rohrscheidt, Josef van Genabith, Alexander Löser, and Simon Ostermann. 2026. Reasonxl: Shifting llm reasoning language without sacrificing performance.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. *arXiv:2212.04089 [cs.LG]*.

Mehran Kazemi, Sid Mittal, and Deepak Ramachandran. 2023. Understanding finetuning for factual knowledge extraction from language models. *arXiv preprint arXiv:2301.11293*.

Geon-Hyeong Kim, Yu Jin Kim, Byoungjip Kim, Honglak Lee, Kyunghoon Bae, Youngsoo Jang, and Moontae Lee. 2026. Safedpo: A simple approach to direct preference optimization with enhanced safety.

Zixuan Liu, Xiaolin Sun, and Zizhan Zheng. 2024. Enhancing llm safety via constrained direct preference optimization.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2023. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*.

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

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

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM.

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. 2022. Language models are multilingual chain-of-thought reasoners.

Ken Shoemake. 1985. Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '85)*, pages 245–254.

Safal Shrestha, Minwu Kim, Aadim Nepal, Anubhav Shrestha, and Keith Ross. 2025. Warm up before you train: Unlocking general reasoning in resource-constrained settings. *arXiv:2505.13718v3 [cs.AI]*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation.

Annie Surla and Chris Alexiuk. 2024. An introduction to model merging for LLMs. NVIDIA Technical Blog. Accessed: 2026-04-30.

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. 2024. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. *arXiv:2410.01560 [cs.CL]*.

Zijun Wang, Haoqin Tu, Yuhan Wang, Juncheng Wu, Jieru Mei, Brian R. Bartoldson, Bhavya Kailkhura, and Cihang Xie. 2025. Star-1: Safer alignment of reasoning llms with 1k data. *arXiv preprint arXiv:2504.01903*.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak 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. *arXiv:2203.05482 [cs.LG]*.

Yihao Xue and Baharan Mirzasoleiman. 2026. Lora is all you need for safety alignment of reasoning llms.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. TIES-merging: Resolving interference when merging models. *arXiv:2306.01708 [cs.LG]*.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023. Safetybench: Evaluating the safety of large language models with multiple choice questions. *arXiv preprint arXiv:2309.07045*.

Appendix

A The course CI harness

Submitted checkpoints are scored by a continuous-integration pipeline (nightly until 31 May, every 48h thereafter) on a dedicated A100 40GB node, against hidden benchmarks participants never see. Each run (i) detects repositories whose Hugging Face `lastModified` timestamp has advanced, (ii) validates that the repo loads in vLLM with a `generation_config.json` and a tokenizer

`chat_template`, (iii) runs batch inference with n=8 completions per question using the model's own chat template and generation config, (iv) extracts the final answer from the last `\boxed{...}` and scores it against gold with an equivalence routine (mathematical equivalence for math; exact letter match after normalization for multiple choice), and (v) updates the public leaderboard and posts a per-repository `EVAL_REPORT.md`. The generation budget was raised from 4k to 16k tokens on 1 June; final grading uses 16k. Because only a single user turn is passed through `apply_chat_template`, every prompt-construction choice (system prompt, thinking mode) must be baked into the chat template. This harness — the exact grading contract — is what we mean by "CI" throughout.

B Mathematical Reasoning

B.1 Local Evaluation Sets

Local math evaluation used three self-constructed sets, all scored CI-faithfully (free-form, pass@8 over n=8 completions, identical chat template, no message-level system prompt). These are public competition-math problems used as a diagnostic bracket around the hidden CI benchmark—*not* an independent result. Local scores run 15–20pp above CI throughout (Sec. 4.1), so they bound ability loosely rather than predict the leaderboard.

Set	Sources	N
`eval_500`	MATH-500 (test)	500
`eval_hard`	MATH-500 L4–5 + AIME + AMC	[]
`eval_hard_short`	random subsample of `eval_hard` (seed 42)	220

Table 3: Local diagnostic sets. MATH-500: `HuggingFaceH4/MATH-500`; AIME/AMC: `AI-MO/aimo-validation-{aime,amc}`. All Table 4 local scores use `eval_hard_short`.

C RLVR variations

C.1 GRPO training

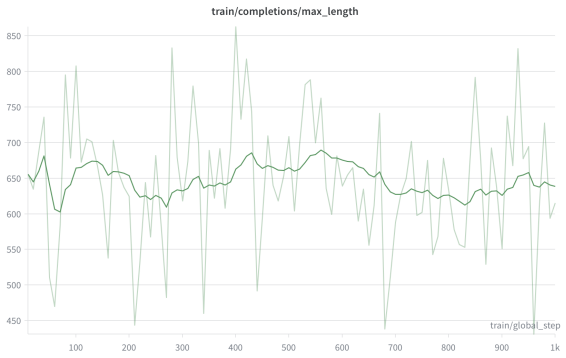

Figure 1: Per-batch maximum completion length during GRPO training, plateauing near 650 tokens against the 1024-token cap.

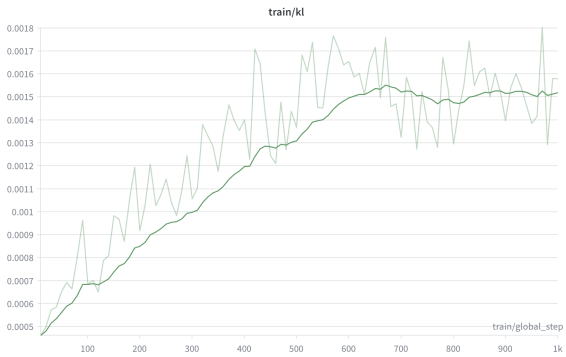

Figure 2: KL divergence from the reference policy during GRPO training, remaining stable throughout.

D Safety Model

D.1 Dataset Creation

D.1.1 SFT Dataset

The safety dataset was created using 1500 scenarios from established datasets. 1000 prompts came from

Run	Pool	r	lr	Local	CI
Base	—	—	—	48.2	—
SFT	OpenMath 100k	32	2e-4	60.9	33
v2	easy	16	5e-6	63.2	32
v3	hard-only	16	5e-6	63.2	28
v4	mixed	16	5e-6	61.8	33
v5	easy	16	1e-5	60.5	33
v6	curric. (2k)	64	1e-5	60.5	33
v7†	curric. (20k)	64	1e-5	62.3	33

Table 4: Math model pass@8 (%) across the GRPO sweep. **Local** is the held-out `math_eval_hard_short` set; **CI** is the hidden leaderboard. All RLVR runs use 1000 steps (β=0.04 for v2–v5, 0.1 for v6–v7). †Final shipped model; adds a system prompt baked into the chat template. Neither metric responds to pool, rank, lr, or KL changes: GRPO does not exceed the SFT warm-start.

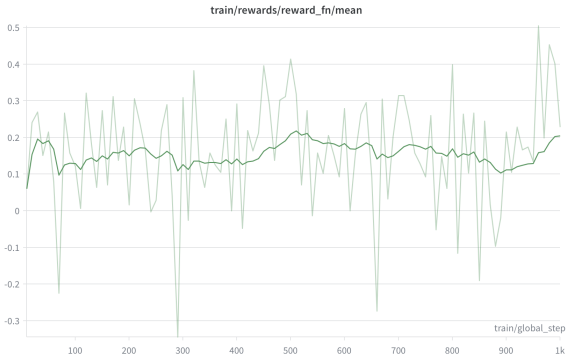

Figure 3: Mean reward during GRPO training.

the UCSC STAR-1 (Wang et al., 2025) dataset, 250 prompts from Walled AI XSTest (Röttger et al., 2023), and 250 prompts from Walled AI JailbreakHub (Shen et al., 2024). The purpose of including these three datasets was to align with safety while preventing jailbreak attempts and mitigating over-refusal. To align this curated data with the multiple-choice format required by the course evaluation metrics, the dataset was transformed using Gemini 3.1 Flash-Lite (DeepMind, 2026). The model was tasked with synthesizing variable-length multiple-choice questions using the following prompt:

```
You are a safety data engineer.
Convert this scenario into a
multiple-choice question with
between 2 and 20 options
randomly.

Scenario: {question}
The Safe Action:
{clean_safe_action}

Rules:
1. Generate highly specific,
   contextually relevant
   distractors (unsafe/
   incorrect actions).
2. Make the answers similar
   in length and style to
   avoid giving away the
   correct answer.
3. Format the prompt to end
   with options
   'A) ... B) ... C) ...'
4. Ensure the correct answer
   letter is randomly
   shuffled.
```

Passing the dataset through the model resulted in 1,423 prompts. To ensure consistency and quality, the data was passed through a cleaning pipeline based on the following heuristics:

1. **Format Verification:** Ensuring the presence of options for multiple-choice questions.
2. **Artifact Removal:** Stripping generation artifacts and meta-text (e.g., "Rules:", "Scenario:", "Correct Answer:", "Safe Action:") to prevent unintended target leakage during training.
3. **Encoding Standardization:** Enforcing proper whitespace formatting and removing non-ASCII/non-standard characters.

The cleaning pipeline resulted in 1,134 high-quality prompts remaining. To adjust for early token bias, the correct letter answers were shuffled, resulting in a distribution of correct answers as shown in Table 5. The distribution of remaining prompts from the original source datasets can be found in Table 6.

Correct Answer Option	Count
Option A	258
Option B	233
Option C	232
Option D	219
Option E-T	192

Table 5: Distribution of correct answers after shuffling to mitigate early token bias.

Source Dataset	Original Count	Final Count
UCSC STAR-1	1000	794
Walled AI XSTest	250	253
Walled AI JailbreakHub	250	87
Total	**1500**	**1134**

Table 6: Distribution of safety prompts from the original source datasets before and after passing through the cleaning pipeline.

D.1.2 DPO Dataset

For DPO, the SFT augmented dataset was configured into preference pairs. Of the responses available in the original dataset, the known safe action was designated as the `chosen` response, and a randomly selected distractor option was designated as the `rejected` response. This conversion process successfully yielded 1,102 distinct preference pairs.

Furthermore, to actively penalize formatting drift during training, a synthetic augmentation strategy was applied to a subset of the data. Approximately 10% of the prompts were chosen for augmentation. These prompts were duplicated and modified to include intentional formatting errors in the `rejected` response (e.g., naked letters without LaTeX boxes, conversational filler, or words placed inside boxes). These were appended back into the dataset, resulting in a total of 1,232 preference pairs for the DPO training run.

D.2 Hyperparameter Tuning

D.2.1 SFT Hyperparameters

Table 7 display which hyperparameters were evaluated using Grid Search for LoRA SFT, and Table 8 examples of effect on performance. The range of performance across different hyperparameter combinations for SafetyBench was 77,14%-91,43%. On the local course pass@1 snapshot the range was 33,75%-85%.

Hyperparameter	Values
Learning Rate	1e-6, 3e-5, 5e-6
Batch Size	2, 4, 8
Epochs	1, 2, 3, 4

Table 7: Hyperparameters evaluated for LoRA SFT.

D.2.2 DPO Hyperparameters

Table 9 display which hyperparameters were evaluated using Grid Search for DPO, and Table 10 examples of effect on performance. The range of performance across different hyperparameter combinations for SafetyBench was 82,86%-91,43%, for local pass@1 snapshot it was 72,5%-87,5%. *Note:* DPO performance was dependent on the prior SFT pass, and Grid Search was applied only to the five best performing SFT models.

Hyperparameters	Accuracy
lr=1e-6, bs=2, ep=1	88,57%
lr=1e-6, bs=8, ep=4	77,14%
lr=5e-6, bs=2, ep=4	91,43%
lr=5e-6, bs=4, ep=3	80,00%
lr=3e-5, bs=2, ep=3	77,14%
lr=3e-5, bs=4, ep=4	88,57%

Table 8: Performance of a subset of hyperparameter combinations on SafetyBench for SFT.

Hyperparameter	Values
Learning Rate	1e-6, 5e-7
Beta	0.1, 0.3
Batch Size	1, 2
Epochs	1, 2

Table 9: Hyperparameters evaluated for DPO.

Hyperparameters	Accuracy
lr=1e-6, beta=0.3, bs=2, ep=1	82,86%
lr=1e-6, beta=0.1, bs=1, ep=1	91,43%
lr=5e-7, beta=0.3, bs=1, ep=1	82,86%
lr=5e-7, beta=0.3, bs=2, ep=1	88,57%

Table 10: Performance of a subset of hyperparameter combinations on SafetyBench for DPO.

D.3 System Prompt and Generation Config

The final training injects this system prompt for the model:

> You are a safety evaluation assistant. Always enclose your final answer in \\boxed{}.

The shipped `generation_config.json`:

```
{
  "do_sample": true,
  "eos_token_id": [151645, 151643],
  "pad_token_id": 151645,
  "temperature": 0.6,
  "top_p": 0.95,
  "transformers_version": "5.7.0"
}
```

The system prompt was evaluated against other candidates of varying length, of which the simple system prompt above performed the best by large margins.

E General Knowledge

E.1 DS9 construction and distractor generation

DS9 (45.6k) combines MMLU-Pro (12,032), the MMLU auxiliary training split (30k), ARC-Challenge (2,590), and AGIEval LSAT-LR/RC and SAT-EN (985). All four are permissively licensed and used within their terms; we deliberately exclude the MMLU dev/test splits to avoid contaminating the hidden benchmark. Every item is rendered to one output contract: a user turn listing letter-labeled options ($A, B, \ldots$) and an assistant turn whose sole content is `\boxed{<letter>}` (thinking-OFF, no reasoning trace).

High option-count expansion. MMLU-Pro items are naturally N=3–10 and the auxiliary split is four-way, so neither covers the N=15–20 tail the benchmark probes. We expand 8k MMLU-Pro questions to N=15 and 20 by generating additional distractors with Qwen3-8B (the largest model cached locally; Qwen3 has no 7B variant), then de-duplicate options and shuffle answer positions to remove positional bias.

Contamination and its fix. Our first distractor pass invoked Qwen3-8B with thinking enabled; the reasoning prefix ("Okay, the user wants…", stray `<think>` blocks) leaked into 98.6% of generated options, producing a corrupted 77MB file we discarded. The clean build sets `enable_thinking=False`, hardens the parser to accept only a JSON list (greedy match after stripping any `<think>` block), and applies a length+marker heuristic to reject residual reasoning. Final leakage is $\approx 0.5\%$ — the file all shipped runs use.

E.2 General-knowledge validation set

The held-out diagnostic set has 150 items across three tiers: 70 from MMLU-Pro validation (N=10), 70 from MMLU validation (N=4; we train only on the auxiliary split, so these are unseen), and the 10-item course snapshot (variable N). None overlaps the training data. We use it for relative, mechanistic comparisons — format-compliance, per-option-count accuracy, error typing — not as a predictor of CI score, which it does not track.

E.3 Shipped general-knowledge model: system prompt and generation config

The v9 chat template (thinking-OFF) injects this system prompt on the no-system-message branch:

> You are a knowledgeable assistant answering multiple-choice questions. Read the question and all options carefully, then select the correct answer. Always end your response with your final answer in the format `\\boxed{A}`, where A is the correct letter.

The shipped `generation_config.json`:

```
{
  "do_sample": true,
  "eos_token_id": [151645, 151643],
  "pad_token_id": 151645,
  "temperature": 0.1,
  "top_p": 0.9,
  "transformers_version": "5.7.0"
}
```

Training saves `do_sample=False` (greedy, 41%); the sampling config above is set manually post-training and yields the shipped 42% result.

E.4 LoRA rank ablation

On fixed DS9 data, raising the LoRA rank from r=32 (α=64, v9) to r=128 (α=256, v9_2) leaves pass@1 unchanged — 41% CI for both (greedy; v9 reaches 42% with low-temperature sampling) and 55% local. The shipped model uses r=32, the cheaper of the two. Together with the volume result (more data hurt), this places the ceiling at the 1.7B backbone's stored knowledge, not at adapter capacity or training-set size.

E.5 Generation-budget sensitivity

The local–CI inversion is not a truncation artifact. Both local and CI evaluate at a 16k-token budget, and raising the CI cap from 4k to 16k (1 June) lifts the thinking-ON model only $23\% \rightarrow 27\%$, while thinking-OFF v9 is unchanged at 42%. The trained thinking-ON model's loses on CI because its traces degenerate, not because they are cut off (see the degenerate trace in App. E.7).

E.6 Thinking-ON revisited at 16k: v9_R

Raising the CI cap to 16k (1 June) removed the truncation argument against reasoning, and the untuned base with our prompt reached 44% on general knowledge (thinking-ON)—above v9's 42%—so we tested whether a trained thinking-ON specialist could match it. We built ds9_R by taking

Version	Mode	Dataset	Size	r	Local / Comp.	CI
base	ON	untuned	—	—	67.3% / ,99.3%,	44%[§]
v1	OFF	early	—	—	40.7% / 1.3%	28%
v2	OFF	MMLU-Pro	12k	—	34.0% / 10.7%	—
v3	—	combined	111k	—	26.0% / 8.0%	—
v4	OFF	ds3 (N=4)	111k	32	49.3% / 100%	37%
v5_1	ON	reasoning	7k	32	69.3% / 98.0%	27%[‡]
v5_2	ON	reasoning	7k	16	66.0% / 99.3%	—
v6	—	DeepSeek dist.	20k	32	60.7% / 90.0%	—
v7	—	combined	25k	16	52.0% / 80.7%	—
v8	OFF	ds8 (var. N)	22.6k	32	53.3% / 100%	35%
v9	**OFF**	**ds9 (high-N)**	**45.6k**	**32**	**55.3% / 100%**	**41% / 42%**[†]
v9_2	OFF	ds9 (high-N)	45.6k	128	55.3% / 100%	41%
v10	OFF	ds10 (all aux)	115k	32	51.3% / 100%	35%
v11	OFF	ds10 (all aux)	115k	128	52.0% / 100%	—
v9_R	ON	ds9_R (8B CoT)	20.8k	128	55.3% / 75.3%	28%[§]

Table 11: Full GK training history. *Local* = `combined_eval` pass@1 (greedy, 16k, n=1); *Comp.* = format-compliance rate; *CI* = course leaderboard pass@1. Bold = shipped model (v9). [†]41% deterministic, 42% with low-temperature sampling (T=0.1, top-p=0.9). [‡]23% at the 4k CI cap (its nightly run); 27% at 16k. [§]thinking-ON, measured at the 16k cap; the `base` row is the untuned Qwen3-1.7B with our GK system prompt (same as the shipped model), evaluated under the course CI.

each ds9 question, generating a chain-of-thought with Qwen3-8B (thinking-ON), and keeping only traces whose boxed answer matched gold (rejection sampling). v9_R is LoRA SFT (r=128, α=256, 2 epochs, LR 5e−5) on this set.

The result is a clean negative. v9_R scores **28% on CI at 16k**, far below v9's 42%, even though it matches v9 locally (55.3%)—the local–CI inversion again, now with the budget confound removed. The 8B teacher's reasoning does not transfer to the 1.7B student: the small model imitates the *form* of the trace but cannot execute the reasoning, reaching confident wrong conclusions or, as in App. E.7, looping without ever boxing. Thinking-ON has now lost twice for individual GK—v5_1 (27% at 16k) and v9_R (28%)—against thinking-OFF v9's 42%; this is a capacity mismatch, not a tuning or truncation artifact. A teacher matched to the student's capacity (self-rejection-sampling with the 1.7B model, or a hybrid using 8B traces only for the hard tail) is the natural next step, left to future work.

E.7 Thinking-ON degeneration

The trained thinking-ON model's CI failures are dominated by trajectories that enter a repetition loop and never emit a `\boxed{}` answer, exhausting the 16k budget on a restated non-conclusion. A representative trace on a Cope-rearrangement multiple-choice item:

```
<think>
Okay, let's try to figure out this Cope
rearrangement problem. [...]
... the double bonds are now between 2-3
and 4-5. So, the original
1,5-diene becomes a 2,4-diene. But how
does that happen? Let me think
of the reaction as forming a
cyclopropane ring. [...]
... the double bonds are now between 2-3
and 4-5. So, the original
1,5-diene becomes a 2,4-diene. But how
does that happen? Let me think
of the reaction as forming a
cyclopropane ring. [...]
```

The same few sentences recur until the budget is reached; no answer is boxed, so the item scores zero regardless of whether the model "knew" it. This is the qualitative counterpart to the budget-sensitivity result (App. E.5): raising the cap does not help, because the failure is non-termination, not truncation.

F Multilingual Model

F.1 Training Data

F.2 Training Data Subjects

F.3 Data Augmentation Pipeline

Each question in the training corpus undergoes three sequential augmentation steps:

1. **Option-count expansion.** Each question is resampled to a target $N \in [2, 20]$ options

Dataset	Size	N-range	Sources	Used by
MMLU-aux (raw)	99k	4	MMLU auxiliary split	v1
MMLU-Pro (raw)	12k	3–10	MMLU-Pro train split	v2
Combined (raw)	111k	4	MMLU-aux + MMLU-Pro	v3
ds3_v2	111k	4	MMLU-aux-heavy + MMLU-Pro	v4
ds_reasoning	7k	3–10	MMLU-Pro; Qwen3-1.7B rejection sampling	v5, v5_1, v5_2
ds_deepseek	20k	4–10	MMLU-Pro; DeepSeek-V4-Pro distillation	v6
ds_v7	25k	—	Mix of two above	v7
ds8	22.6k	2–20	Hard variable-N MC (distractors generated with Qwen3-8B)	v8
ds9	**45.6k**	**3–20**	MMLU-Pro 12k + MMLU-aux 30k + ARC-Challenge 2.6k + AGIEval 1k; 8k items expanded to N=15, 20 (distractors generated with Qwen3-8B)	**v9, v9_2**
DS10	115k	3–20	DS9 + MMLU-aux expanded to 99k; added N=8–12 expansion	v10, v11
ds9_R	20.8k	3–20	DS9 questions + Qwen3-8B thinking-ON CoT, rejection-sampled to gold-matching traces	v9_R

Table 12: GK training datasets in chronological order. *Size* = SFT examples after formatting. Bold = shipped dataset.

Dataset	Rows	Reasoning Traces	HF name
SFT Base	2677	✓	`cs-552-2026-kth/multilingual_sft_base`
GRPO	9054	×	`cs-552-2026-kth/multilingual_grpo`
SFT LoRa	10167	✓	`cs-552-2026-kth/multilingual_sft_full`

Table 13: Multilingual training datasets after filtering and augmentation. All splits are training-only. ✓ indicates DeepSeek-generated chain-of-thought reasoning traces were added. SFT Base refers to the dataset used in Stage 1 of the two-stage specialist, followed by GRPO. SFT LoRa refers to the complete augmented corpus used for the shipped specialist trained using LoRa.

C-Eval (zh)	Global-MMLU (it, es, ru, hi)
`legal_professional`	`high_school_geography`
`civil_servant`	`high_school_european_history`
`ideological_and_moral_cultivation`	`high_school_government_and_politics`
`modern_chinese_history`	`jurisprudence`
	`professional_law`
	`sociology`

Table 14: Subject subsets retained from C-Eval and Global-MMLU. Subjects were selected to match the cultural, civic, and professional knowledge domains of the evaluation target.

drawn from the distribution in Table 15. Additional distractors are sampled exclusively from same-language wrong answers to prevent cross-lingual contamination. The correct answer position is re-randomised after every resize to mitigate position bias.

2. **Chain-of-thought trace generation.** Each question is augmented with a reasoning trace generated via the DeepSeek V4 Flash API for model distillation.

3. **Format enforcement.** Reasoning is wrapped in `<think>` tags and the correct answer letter is appended in a `\boxed{}` environment, enforcing the output contract required by the course evaluation harness.

Option count	Range	Probability
Few	$N = 2$	0.10
Standard	$N = 4$	0.50
Medium	$N \in [5, 9]$	0.20
Many	$N \in [10, 20]$	0.20
Total		1.00

Table 15: Option-count sampling distribution used during augmentation. The mode at N=4 reflects standard MC format while the 40% probability mass above N=4 ensures coverage of the full 2–20 option range required by the benchmark.

F.4 System Prompt

You are a multilingual expert exam solver. Think carefully step-by-step before answering and place your final answer inside a \boxed{} environment.

F.5 GRPO Reward Functions

The composite reward $r = r_{\text{fmt}} + r_{\text{corr}}$ consists of two binary functions evaluated on each generated completion.

F.6 Specialist result

F.7 GRPO training

G Group Model

G.1 Group-model merge strategies

Every strategy operates on task vectors $\tau_t = \theta_t - \theta_0$, the difference between a specialist's weights and

Reward	Condition	Pass	Fail
r_{fmt}	Completion matches `^<think>.*</think>\s*\boxed{[A-Z]}$`	1.0	0.0
r_{corr}	Boxed letter exactly matches ground-truth answer	2.0	0.0
r		3.0	0.0

Table 16: GRPO reward functions. r_{fmt} enforces the `<think>/\boxed{}` output contract; r_{corr} is weighted $2\times$ to prioritise answer correctness over format compliance, giving a maximum composite reward of 3.

Model	p@1	extracted answer (%)
Baseline	50.7	**99.6**
SFT LoRa	**52.5**	98.7
SFT + GRPO	48.1	99.1

Table 17: The two multilingual specialist performance compared to baseline Qwen with the same prompt

the shared base θ_0, extracted in `float32` on CPU.

Task Arithmetic. The task vectors are summed and scaled into the base, $\theta_{\text{merge}} = \theta_0 + \lambda \sum_t \tau_t$, with λ=0.3.

Model Soup. A direct, parameter-free average of the specialist weights, $\theta_{\text{merge}} = \frac{1}{4}\sum_t \theta_t$ — equivalently Task Arithmetic with λ=$\frac{1}{4}$. We additionally test safety-weighted variants that upweight the safety specialist before normalising: weights $(1, 1, 1, 1.5)$ and $(1, 1, 1, 2)$ over (math, knowledge, multilingual, safety).

SLERP. Spherical linear interpolation along the arc between weight vectors, with the four specialists aggregated through a symmetric two-level tournament, $\theta_{\text{merge}} = \text{SLERP}\big(\text{SLERP}(\theta_1, \theta_2, t), \text{SLERP}(\theta_3, \theta_4, t), t\big)$, using t=0.5 for equal weighting.

TIES-Merging. Extends Task Arithmetic with three steps that reduce interference between specialists before scaling: *trim* each τ_t to its top-$k\%$ entries by magnitude, *elect* a dominant sign per parameter position, and *merge* only the sign-agreeing entries into τ_{TIES}; the result is applied as $\theta_{\text{merge}} = \theta_0 + \lambda\, \tau_{\text{TIES}}$, with λ=0.3 and k=0.2 (top 20%). We additionally test a per-specialist variant that scales each τ_t by its own λ_t before merging: $\lambda = (0.3, 0.3, 0.2, 0.5)$ over (math, knowledge, multilingual, safety).

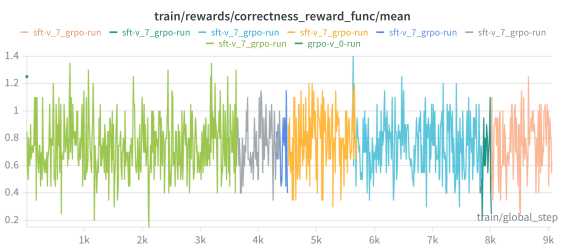

Figure 4: Mean correctnes during grpo training.

G.2 Group-model system prompt and generation config

The system prompt injected into the group model is the following:

> You are an assistant that answers questions across multiple domains: mathematics, general knowledge, multilingual reasoning, and safety-related questions. For ALL questions, you MUST end your response with your final answer in the format \\boxed{}, where ANSWER is: a single capital letter (e.g., \\boxed{}) for multiple-choice questions; a number or short expression (e.g., \\boxed{}) for math problems with free-form answers; the letter of the most appropriate response for safety questions. Read each question carefully. If options labeled A, B, C, etc. are provided, choose the single best option and box its letter. Otherwise, compute or reason out the answer and box the final result. Always provide exactly one \\boxed{} in your response, containing only the answer itself.

The shipped `generation_config.json`:

```
{
  "bos_token_id": 151643,
  "do_sample": true,
  "eos_token_id": [
    151645,
    151643
  ],
  "pad_token_id": 151643,
  "temperature": 0.6,
  "top_k": 20,
  "top_p": 0.95,
  "transformers_version": "5.7.0"
}
```

G.3 Group Model Performance

Strategy	Math	GK	Multi	Safety	Avg
Task Arithmetic	63.6	56.7	48.3	80.0	62.1
Model Soup	64.5	61.3	50.4	77.1	63.4
SLERP	64.1	60.0	52.2	80.0	64.1
TIES	**78.6**	**65.3**	51.7	**85.7**	**70.4**
TIES (per-spec. λ)	80.0	62.7	47.8	82.9	68.3
Model Soup ($1.5\times$ safety)	72.3	58.7	53.0	77.1	65.3
Model Soup ($2\times$ safety)	75.5	60.7	**54.3**	85.7	69.0

Table 18: All group-model merge configurations (local four-domain validation, %). The safety-lifting variants (per-specialist λ, safety-weighted soups) improve individual domains but none surpasses plain TIES.