UNCOVERING CROSS-OBJECTIVE INTERFERENCE IN MULTI-OBJECTIVE ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

We study a persistent failure mode in multi-objective alignment for large language models (LLMs), in which scalarized training improves only a subset of objectives while the others degrade. We formalize this phenomenon as *cross-objective interference* and conduct, to our knowledge, the first systematic study of nine scalarization algorithms, finding it pervasive yet strongly model-dependent. To explain interference, we derive a local covariance law in which an objective rises or falls with the sign of its covariance with the scalarized score. We extend this analysis to clipped surrogate objectives used in modern alignment, demonstrating that the covariance law remains valid under mild conditions despite clipping. Building on this analysis, we propose COVariance-floor Enforced Reweighting (COVER), a plug-and-play controller that raises an objective's weight only when its reward's covariance with the clipped advantage weight falls below a target. Through extensive experiments, we show that COVER effectively mitigates cross-objective interference while matching linear scalarization when objectives already co-improve. Finally, we complement these local improvement conditions with a global convergence analysis that establishes when non-convex scalarized optimization satisfies the Polyak–Łojasiewicz condition and how cross-objective interference depends on model geometry.[1]

1 INTRODUCTION

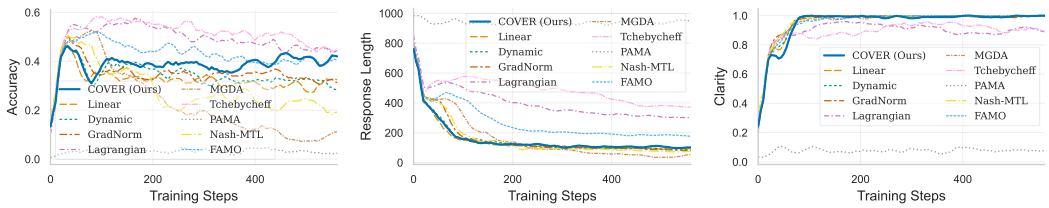
Multi-objective LLM alignment predominantly builds on reinforcement fine-tuning (RFT) (Ouyang et al., 2022; Bai et al., 2022; Lambert et al., 2025; Shen et al., 2025), commonly reducing the problem to a single scalar objective through *scalarization* with static (Kimi et al., 2025) or dynamic (Lu et al., 2026) weights at the reward (Guo et al., 2024) or gradient level (Li et al., 2025). Despite its simplicity and popularity, scalarization has a persistent and underexplored failure mode. It frequently fails to improve all objectives simultaneously, and training keeps improving a subset of "easy" objectives while others degrade, a pattern we formalize as *cross-objective interference*.

To investigate whether this is an artifact of naive implementations or a fundamental limitation, we evaluate nine well-established Multi-Task Learning (MTL) and Multi-Objective Optimization (MOO) algorithms at both scalarization levels. Reward-level methods include classical linear weighting (Barrett & Narayanan, 2008) and Lagrangian primal-dual (Mahdavi et al., 2013), and more recent smooth Tchebycheff (Tchebycheff hereafter, Lin et al., 2024; Li et al., 2026), PAMA (He & Maghsudi, 2026), and dynamic reward weighting (Lu et al., 2026). Gradient-level methods aggregate per-objective gradients into a single update direction, as in MGDA (Désidéri, 2009), GradNorm (Chen et al., 2018), Nash-MTL (Navon et al., 2022), and FAMO (Liu et al., 2023). To our knowledge, this is the first systematic evaluation of scalarization algorithms for multi-objective LLM alignment.

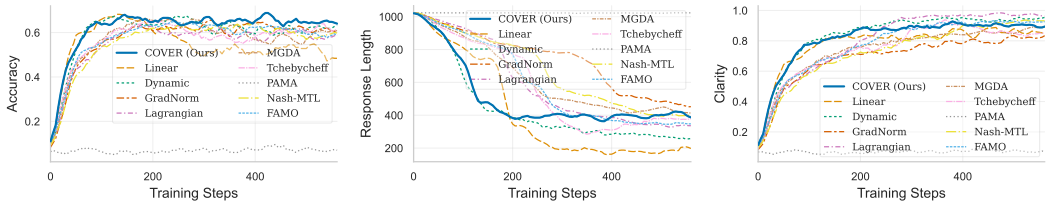
Our results reveal that interference is pervasive yet strongly model-dependent. On Qwen2.5-1.5B-Base (Figure 1a) and Qwen2.5-1.5B-IFT, accuracy falls from its peak while the auxiliary objectives improve in 18 of 20 runs. Under the same training procedure, however, Qwen3-1.7B-Base exhibits interference in only 1 of 10 runs (Figure 1b). Beyond these small models, the effect also appears at 7B and 8B parameters across model families and datasets. Critically, it arises even when the objectives do not strongly conflict under the gradient-based definitions commonly used in MTL and

[1]We open-source code to reproduce our results: anonymous.4open.science/r/COVER-anonymous-BB1C.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Qwen2.5-1.5B-Base



(b) Qwen3-1.7B-Base

Figure 1: **MATH-500 test performance over training on Qwen2.5-1.5B-Base and Qwen3-1.7B-Base.** Test accuracy, response length (lower is more concise), and clarity for nine scalarization algorithms and COVER. On Qwen2.5-1.5B-Base, every method except COVER (ours) and PAMA (which fails to learn) improves conciseness and clarity at the expense of accuracy, a clear instance of cross-objective interference. Under the same training procedure, eight of ten methods co-improve all three objectives on Qwen3-1.7B-Base, only Linear interferes, and PAMA again fails to learn.

MOO (Evgeniou & Pontil, 2004; Yu et al., 2020; Liu et al., 2021; SHI et al., 2023; Kim et al., 2025).[2] Existing theories on linear scalarization (Lu et al., 2023), convexity (Wei & Niethammer, 2021), gradient conflict (Sener & Koltun, 2018), and generalization tradeoffs (Chen et al., 2023) leave these findings unexplained, because they were developed for simplified settings rather than LLM alignment.

To address this gap, we develop a theoretical framework analyzing multi-objective alignment through first-order improvement conditions. Starting from the classic policy gradient (Sutton et al., 1999; Hu et al., 2025), we derive a reward-level local covariance law that precisely characterizes when an objective improves under scalarized alignment, namely *when its true reward exhibits positive covariance with the scalarized score*. This explains cross-objective interference. Objectives that are easy to optimize can dominate training, inducing negative covariance for harder objectives that then degrade even as the overall scalarized return increases.

We then extend this analysis to clipped surrogate objectives used in modern RFT, such as GRPO (Shao et al., 2024), showing that the covariance law remains valid under mild conditions despite clipping. This directly motivates our method, COVariance-floor Enforced Reweighting (COVER), a one-sided controller that raises an objective's weight only when the covariance between its reward and the clipped training signal falls below a target. When every covariance reaches its target, COVER leaves the weights unchanged and trains like linear scalarization.

Being local, the covariance law predicts each objective's next change, not where training converges, so it cannot explain the strong model dependence we observe (e.g., Qwen3-1.7B-Base). To address this fundamental problem, we study the global geometry of scalarized RFT using the Polyak–Łojasiewicz (PL) inequality, which accommodates non-convex objectives. We derive sufficient conditions under which the scalarized RFT objective satisfies a μ-PL inequality, yielding a concrete, model-dependent mechanism for when cross-objective interference arises: (i) the policy assigns insufficient probability mass to the optimal trajectory, (ii) the scalarization yields weak reward margins between optimal and suboptimal trajectories, or (iii) token-level gradient contributions cancel under an ill-conditioned Jacobian. Together, these analyses suggest that cross-objective interference is both algorithmic (covariance misalignment) and model-dependent (unfavorable geometry), and provide actionable insights for robust multi-objective LLM alignment. In summary, our contributions are threefold:

[2]See Appendix A, Figures 4 to 6 for results across models and scales, and Figure 7 for gradient analysis.

- **Systematic empirical study:** We provide the first systematic evaluation of MTL and MOO scalarization algorithms for LLM alignment, finding that cross-objective interference is pervasive and strongly model-dependent (§4.4).
- **Local improvement theory and method:** We derive a reward-level local covariance law characterizing first-order conditions for objective improvement (§4.1), extend it to clipped surrogate objectives (§4.3), and propose COVER to mitigate cross-objective interference (§5).
- **Global convergence analysis:** We derive sufficient conditions for scalarized RFT to satisfy the PL inequality and link cross-objective interference to model geometric properties, thus laying theoretical foundations for future work (§6).

2 Related Work

2.1 Multi-Task Learning: Gradient Conflicts and Solutions

MTL addresses joint training across multiple losses, where negative transfer often arises from conflicting gradients and imbalanced loss scales. Common solutions either reweight tasks, as in GradNorm (Chen et al., 2018), Nash-MTL (Navon et al., 2022), and FAMO (Liu et al., 2023), or modify gradients directly, as in PCGrad (Yu et al., 2020), CAGrad (Liu et al., 2021), Gradient Vaccine (Wang et al., 2021), Recon (SHI et al., 2023), and DB-MTL (Lin et al., 2025). While these MTL methods provide valuable insights for gradient-level control or loss balance, they are developed primarily for supervised settings with convex or well-behaved loss structures.

2.2 Multi-Objective Optimization: Scalarization and Pareto Optimality

Classic MOO reduces multi-objective problems to single-objective optimization via scalarization, with convergence guarantees typically relying on convexity or determinism assumptions that do not transfer cleanly to LLM alignment. The foundational gradient-based approach MGDA (Désidéri, 2009) computes a common descent direction by solving a minimum-norm problem over the convex hull of objective gradients. Recent extensions include PMGDA (Zhang et al., 2024), which incorporates user preferences, and PAMA (He & Maghsudi, 2026), which solves a reward-level minimum-norm problem for LLM alignment. While Lu et al. (2023) analyze when linear scalarization can in principle recover the full (non-convex) Pareto front, their guarantees rely on strong non-determinism and numerical stability assumptions. Stochastic MOO methods such as SMG (Liu & Vicente, 2021), MoCo (Fernando et al., 2023), and SDMGrad (Xiao et al., 2023) relax deterministic assumptions but remain primarily designed for supervised settings.

Recent work has refined Pareto stationarity concepts (Hu & Yu, 2025) and explored regimes where multiple objectives can facilitate optimization (Dann et al., 2023; Efroni et al., 2025). In multi-objective RL, CA-NPG (Gu et al., 2025) and conflict-averse updates (Kim et al., 2025) aim to improve all objectives under KL or safety constraints, but they do not address the covariance misalignment we identify. None of these works formalizes cross-objective interference or explains its dependence on model geometry.

2.3 Multi-Objective LLM Alignment

Most multi-objective alignment studies reduce multiple rewards to a scalar objective through reward-level scalarization, including static linear scalarization (Wu et al., 2023; Zhang & Zuo, 2025; Yao et al., 2025), dynamic weighting (Lu et al., 2026), Lagrangian relaxation (Moskovitz et al., 2024), Tchebycheff scalarization and its variants (Steuer & Choo, 1983; Lin et al., 2024; Li et al., 2026), or reward-decoupled advantage normalization (Liu et al., 2026). These methods can be extended to produce steerable policies that adapt to user preferences (Basaklar et al., 2023; Wang et al., 2024; Xie et al., 2025). Gradient-level scalarization builds update directions in parameter space, such as GAPO (Li et al., 2025), which requires one backward pass per objective. COVER also adapts reward-level weights, but it reacts only to a per-objective covariance signal and needs no per-objective gradients.

2.4 Challenges in Reinforcement Fine-tuning

LLM RFT faces several optimization challenges beyond multi-objective settings. Lagrangian dynamics can become unstable when convexity assumptions fail (Feijer & Paganini, 2010). LLM-specific challenges include vanishing gradients (Razin et al., 2024), sensitivity to importance weighting and normalization (Zheng et al., 2025), and exploration difficulties (Jiang et al., 2025). Recent RL scaling laws further suggest that optimization performance varies with model size (Khatri et al., 2025). Closest to our setting is gradient starvation (Pezeshki et al., 2021), where a dominant feature captures the gradient at the expense of others. However, its solution, spectral decoupling, collapsed the policy at every strength we tried, even when training on accuracy alone (Appendix Table 10).

3 Preliminaries

Notation. Following the notations from Razin et al. (2024), let $\mathcal{D}$ be the dataset and $\mathcal{X}$ be a finite token vocabulary. We then define $\mathcal{X}^{L_{\text{in}}}$ as the space of input prompts of length L_{in}, and $\mathcal{X}^{L_{\text{out}}}$ the space of output sequences of length L_{out}. We study M objectives and for a given input prompt $\mathbf{x} = (x_1, x_2, \cdots, x_{L_{\text{in}}}) \in \mathcal{X}^{L_{\text{in}}}$ and generated completion $\mathbf{y} = (y_1, \ldots, y_{L_{\text{out}}}) \in \mathcal{X}^{L_{\text{out}}}$, the reward function is $r : \mathcal{X}^{L_{\text{in}}} \times \mathcal{X}^{L_{\text{out}}} \to \mathbb{R}^M$.

RFT as a contextual bandit. We model RFT as a horizon-one (bandit) environment: each input is a state and each completion an action. An autoregressive language model with parameters $\theta \in \mathbb{R}^n$ induces $p_\theta(\mathbf{y} \mid \mathbf{x}) = \prod_{l=1}^{L_{\text{out}}} p_\theta\big(y_l \mid \mathbf{x}, \mathbf{y}_{\le l-1}\big) = \prod_{l=1}^{L_{\text{out}}} \operatorname{softmax}\big(f(\mathbf{x}, \mathbf{y}_{\le l-1}; \theta)\big)_{y_l}$, where $\mathbf{y}_{\le l-1} \coloneqq (y_1, \cdots, y_{l-1})$ is the partial completion and $f(\mathbf{x}, \mathbf{y}_{\le l-1}; \theta) \in \mathbb{R}^{|\mathcal{X}|}$ is the next-token logit map. For each objective $m \in \{1, \ldots, M\}$, the expected objective reward is $r_m(p_\theta) \coloneqq \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\, \mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[r_m(\mathbf{x}, \mathbf{y})\big]$.

Scalarization. A scalarization map $\Psi : \mathbb{R}^M \to \mathbb{R}$ converts the vector reward to a per-sample score $s(\mathbf{x}, \mathbf{y}) \coloneqq \Psi(r(\mathbf{x}, \mathbf{y}))$. The induced value function for input $\mathbf{x}$ is

$$V(\mathbf{x}; \theta) \coloneqq \mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[\, s(\mathbf{x}, \mathbf{y})\,\big], \tag{1}$$

and the overall RFT objective is to maximize $V(\theta) \coloneqq \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\big[V(\mathbf{x}; \theta)\big]$.

4 Local Covariance Laws for Multi-Objective Policy Improvement

In this section, we establish sufficient conditions under which optimizing a scalarized score $s(\mathbf{x}, \mathbf{y})$ guarantees first-order improvement in objectives $r_m(\mathbf{x}, \mathbf{y})$. We analyze a KL-regularized improvement step in distribution space (§4.1), instantiate it in a closed-form two-mode example where scalarization provably hurts an objective (§4.2), extend it to clipped surrogate objectives (§4.3), and evaluate it empirically (§4.4). Proofs are deferred to Appendix D.

4.1 KL-Regularized Policy Improvement

For fixed $\mathbf{x}$, write $p_{\theta;\mathbf{x}}(\cdot) \coloneqq p_\theta(\cdot \mid \mathbf{x})$ and define the KL-regularized improvement step in distribution space:

$$p^+_{\theta;\mathbf{x}} \coloneqq \arg\max_{q \in \Delta(\mathcal{X}^{L_{\text{out}}})} \Big\{ \mathbb{E}_{\mathbf{y}\sim q}\big[s(\mathbf{x}, \mathbf{y})\big] - \frac{1}{\eta}\, \mathrm{KL}\big(q \| p_{\theta;\mathbf{x}}\big) \Big\}, \tag{2}$$

where $\eta > 0$ is the step size and $\Delta(\mathcal{X}^{L_{\text{out}}})$ is the simplex over completions. Its optimizer is an exponential tilt of the current policy, $p^+_{\theta;\mathbf{x}}(\mathbf{y}) \propto p_{\theta;\mathbf{x}}(\mathbf{y}) \exp\big(\eta\, s(\mathbf{x}, \mathbf{y})\big)$ (Theorem D.1, Appendix D).

Theorem 4.1 (First-order local covariance law)**.** *Assume $r_m(\mathbf{x}, \mathbf{y})$ is bounded and there exists $\eta_0 > 0$ such that, for all $\mathbf{x} \in \mathcal{X}^{L_{\text{in}}}$ and $|\eta| \le \eta_0$, $\mathbb{E}_{\mathbf{y}\sim p_{\theta;\mathbf{x}}}\big[\exp(\eta\, s(\mathbf{x}, \mathbf{y}))\big] < \infty$. Then for each objective $m \in \{1, \ldots, M\}$,*

$$r_m(p^+_\theta) - r_m(p_\theta) = \eta \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\Big[\mathrm{Cov}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big(r_m(\mathbf{x}, \mathbf{y}),\, s(\mathbf{x}, \mathbf{y})\big)\Big] + O(\eta^2). \tag{3}$$

Consequently, if $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\Big[\mathrm{Cov}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big(r_m(\mathbf{x}, \mathbf{y}), s(\mathbf{x}, \mathbf{y})\big)\Big] > 0$, then $r_m(p^+_\theta) > r_m(p_\theta)$ for sufficiently small $\eta > 0$.

Equation (3) shows that optimizing the scalar score s improves objective m at first order when completions with higher $s(\mathbf{x}, \mathbf{y})$ also tend to have higher $r_m(\mathbf{x}, \mathbf{y})$, leading to positive covariance averaged across prompts. Conversely, negative covariance brings a local tradeoff where increasing s necessarily decreases r_m at first order.

Because covariance is computed on-policy, its sign can flip over training as p_θ moves. The scalarized update may improve objective m early in training but degrade it later, even when objectives are not inherently conflicting on the global Pareto front. For linear scalarization $s_\lambda = \sum_j \lambda_j r_j$, the condition becomes $\mathrm{Cov}(r_m, s_\lambda) = \sum_j \lambda_j \mathrm{Cov}(r_m, r_j)$, making cross-objective interference more concrete: emphasizing an easy objective can flip the sign for a harder one when their rewards are negatively correlated on-policy.

4.2 A Two-Mode Toy Example: When Scalarization Hurts an Objective

We analyze a minimal setting where the completion distribution places most of its mass on two modes (e.g., two distinct styles or solutions that the model frequently samples). Formally, for a fixed prompt $\mathbf{x}$, we idealize the completion space by two canonical outputs $\mathcal{Y}(\mathbf{x}) = \{\mathbf{y}_{\mathrm{good}}, \mathbf{y}_{\mathrm{bad}}\}$, where "good" and "bad" are defined with respect to a particular objective m, not the scalar score. Let $p_t \coloneqq p_{\theta_t;\mathbf{x}}(\mathbf{y}_{\mathrm{bad}})$ so that $1 - p_t = p_{\theta_t;\mathbf{x}}(\mathbf{y}_{\mathrm{good}})$ and define

$$s_{\mathrm{good}} \coloneqq s(\mathbf{x}, \mathbf{y}_{\mathrm{good}}),\ s_{\mathrm{bad}} \coloneqq s(\mathbf{x}, \mathbf{y}_{\mathrm{bad}}),\ r_{\mathrm{good}} \coloneqq r_m(\mathbf{x}, \mathbf{y}_{\mathrm{good}}),\ r_{\mathrm{bad}} \coloneqq r_m(\mathbf{x}, \mathbf{y}_{\mathrm{bad}}).$$

Optimizing s concentrates probability on what s favors. By Theorem D.1, after one KL-regularized improvement step, the probability of $\mathbf{y}_{\mathrm{bad}}$ becomes $p_{t+1} = (p_t\, e^{\eta s_{\mathrm{bad}}})/(p_t\, e^{\eta s_{\mathrm{bad}}} + (1 - p_t)\, e^{\eta s_{\mathrm{good}}})$, and the log-odds update is

$$\log \frac{p_{t+1}}{1 - p_{t+1}} = \log \frac{p_t}{1 - p_t} + \eta\big(s_{\mathrm{bad}} - s_{\mathrm{good}}\big). \tag{4}$$

If the proxy score ranks the bad mode higher, $s_{\mathrm{bad}} > s_{\mathrm{good}}$, then the log-odds increase linearly in t, and $p_t \to 1$ (i.e., the update drives mass toward $\mathbf{y}_{\mathrm{bad}}$).

Biased objective can decrease monotonically over steps. Assume the objective m prefers the $\mathbf{y}_{\mathrm{good}}$ mode, i.e., $r_{\mathrm{good}} > r_{\mathrm{bad}}$. Then the expected objective for input $\mathbf{x}$ is

$$\mathbb{E}_{\mathbf{y}\sim p_{\theta_t}(\cdot|\mathbf{x})}\big[r_m(\mathbf{x}, \mathbf{y})\big] = (1 - p_t)\, r_{\mathrm{good}} + p_t\, r_{\mathrm{bad}} = r_{\mathrm{good}} - p_t\, (r_{\mathrm{good}} - r_{\mathrm{bad}}),$$

which is strictly decreasing in p_t. Combining with Equation (4) yields a simple cross-objective interference: if the scalarized score favors the mode that is worse for objective m ($s_{\mathrm{bad}} > s_{\mathrm{good}}$) while objective m prefers the other mode ($r_{\mathrm{good}} > r_{\mathrm{bad}}$), then training increases p_t at each step and $\mathbb{E}[r_m(\mathbf{x}, \mathbf{y})]$ decreases monotonically toward r_{bad}.

The covariance law flags the failure. In this two-mode example, the conditional covariance has a closed form:

$$\mathrm{Cov}_{\mathbf{y}\sim p_{\theta_t}(\cdot|\mathbf{x})}\big(r_m(\mathbf{x}, \mathbf{y}), s(\mathbf{x}, \mathbf{y})\big) = p_t(1 - p_t)\, (r_{\mathrm{good}} - r_{\mathrm{bad}})\, (s_{\mathrm{good}} - s_{\mathrm{bad}}). \tag{5}$$

Under the interference configuration $s_{\mathrm{bad}} > s_{\mathrm{good}}$ and $r_{\mathrm{good}} > r_{\mathrm{bad}}$, Equation (5) is negative when $p_t \in (0, 1)$. Therefore, our local covariance law predicts that, for sufficiently small η, the KL-regularized improvement step that increases s must decrease the true objective r_m at first order.

4.3 Extension to Clipped Surrogate Objectives

The two-mode example is illustrative, but modern RFT methods such as GRPO and PPO (Schulman et al., 2017) optimize clipped surrogate objectives rather than the idealized value function. In this section, we show that the derived covariance law continues to hold under mild conditions. The key observation is that clipping does not introduce a fundamentally different improvement mechanism, and it only removes a subset of weighted logit-gradient terms from the update. Therefore, when the removed gradient mass is sufficiently small, the first-order improvement guarantee is preserved. For simplicity, we defer the full formal analysis to Appendix C and present the simplified corollary below.

Table 1: **Per-objective covariance and test-performance change** on MATH-500 and IFEval. *Mean covariance*: $\mathbb{E}_{t,\mathbf{x}}\big[\mathrm{Cov}_{\mathbf{y}}(r_m, s_{\lambda_t})\big]$ over training steps t and prompts $\mathbf{x}$, with λ_t the method's weights at step t. *Negative fraction*: fraction of training prompts with a negative covariance. *Test performance* Δ_m: mean of the last five evaluations minus initialization. †Conciseness starts at its ceiling of 1.0, so $\Delta_m \leq 0$ by design (Appendix A.1). Under all three methods, helpfulness has the smallest mean covariance and the largest negative fraction, and is the only objective besides conciseness to decline.

		Mean covariance ($\times 10^{-2}$)			Negative fraction			Test performance Δ_m		
Dataset	Objective	Linear	Dynamic	COVER	Linear	Dynamic	COVER	Linear	Dynamic	COVER
MATH-500	Accuracy	2.40 ± 0.09	2.40 ± 0.08	2.70 ± 0.16	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	$+0.34 \pm 0.05$	$+0.34 \pm 0.02$	$+0.33 \pm 0.02$
	Clarity	0.91 ± 0.14	0.83 ± 0.05	0.93 ± 0.05	0.001 ± 0.000	0.001 ± 0.000	0.001 ± 0.000	$+0.96 \pm 0.01$	$+0.96 \pm 0.01$	$+0.97 \pm 0.00$
	Conciseness†	1.40 ± 0.20	1.34 ± 0.05	1.49 ± 0.09	0.002 ± 0.000	0.003 ± 0.000	0.004 ± 0.001	-0.03 ± 0.01	-0.05 ± 0.01	-0.06 ± 0.02
IFEval	Constraint following	1.72 ± 0.02	1.75 ± 0.03	1.81 ± 0.01	0.005 ± 0.001	0.004 ± 0.001	0.004 ± 0.001	$+0.19 \pm 0.04$	$+0.20 \pm 0.03$	$+0.21 \pm 0.02$
	Helpfulness	0.03 ± 0.01	0.04 ± 0.03	0.06 ± 0.02	0.416 ± 0.003	0.420 ± 0.017	0.419 ± 0.007	-0.04 ± 0.03	-0.04 ± 0.02	-0.03 ± 0.03
	Conciseness†	1.75 ± 0.03	1.81 ± 0.01	1.84 ± 0.03	0.006 ± 0.000	0.006 ± 0.001	0.007 ± 0.001	-0.10 ± 0.03	-0.10 ± 0.03	-0.11 ± 0.02

Corollary 4.2 (Clipping robustness (Simplified))**.** *Suppose the unclipped surrogate update has a positive first-order improvement margin for objective m (its direction locally increases r_m), and let the clipping distortion be the Fisher-weighted gradient mass removed by clipping. If the distortion is smaller than the margin, the clipped update still yields $r_m(\theta^+) \geq r_m(\theta)$ for small enough $\eta > 0$.*

We also find that in the categorical bandit case, this condition reduces to the same covariance form as Theorem 4.1 (see Corollary C.3). Importantly, this clipping distortion can be carefully controlled in practice through techniques such as learning rate scheduling or reward normalization, keeping clipping distortion small enough for the covariance law to hold. We empirically validate this in Figure 8, where GRPO yields results consistent with the clipping-free REINFORCE findings reported in Figure 1, demonstrating the robustness of our covariance law.

4.4 Experiments: Cross-Objective Interference and Covariance Law

We first describe the experimental setup shared across the paper, then show that cross-objective interference is pervasive and that the covariance sign helps identify which objective degrades.

4.4.1 Setup

Domains, objectives, and rewards. We use six datasets: MATH-500 (Hendrycks et al., 2021; Lightman et al., 2024) and GSM8K (Cobbe et al., 2021) for math, IFEval (Zhou et al., 2023) for instruction following, MBPP (Austin et al., 2021) for code, PKU-SafeRLHF (Ji et al., 2025) for safety, and XSum (Narayan et al., 2018) for summarization. Each pairs a task objective with auxiliary objectives such as clarity and conciseness (Appendix A.1).

Training and evaluation. Each dataset is split 0.6/0.2/0.2 into train/validation/test. We evaluate on the test set every 5 training steps and report only test results. We repeat each run with three seeds and report the mean and standard deviation. Full hyperparameter settings are provided in Appendix Table 11.

Models and baselines. We train Qwen2.5-1.5B-Base, Qwen2.5-1.5B-IFT (Qwen et al., 2025), Qwen3-1.7B-Base (Yang et al., 2025), and SmolLM2-1.7B-Base (allal et al., 2025), alongside Qwen2.5-7B-Base (Qwen et al., 2025), DeepSeek-7B-Chat (DeepSeek-AI et al., 2024), Llama3.1-8B-IFT (Grattafiori et al., 2024), Ministral-8B-IFT (Mistral AI Team, 2024), and Qwen3-8B-Base (Yang et al., 2025) at larger scale. We compare against nine baselines: five reward-level (Linear, Dynamic, Lagrangian, Tchebycheff, PAMA) and four gradient-level (MGDA, GradNorm, Nash-MTL, FAMO). Appendix E gives pseudocode for baselines adapted to LLM alignment. All main experiments use unclipped REINFORCE++ (Hu et al., 2025), and we rerun MATH-500 under clipped GRPO.

4.4.2 Results and Analysis

Interference is pervasive. Under Theorem A.1, 19 of the 30 REINFORCE runs on MATH-500 interfere, 18 of them on the two Qwen2.5 models, and accuracy is always the degraded objective.

Domains. Under Linear, MBPP pass rate falls ($0.565 \rightarrow 0.424$) and XSum coverage collapses ($0.241 \rightarrow 0.054$) as summaries shrink, while GSM8K and PKU-SafeRLHF show no interference (Appendix Table 8). **Scale.** At 7B and 8B, accuracy again falls from its peak on Qwen2.5-7B-Base ($0.67 \rightarrow 0.51$ under Dynamic) and Ministral-8B-IFT while clarity saturates, whereas Qwen3-8B-Base co-improves like its 1.7B counterpart (Appendix Figure 6). **Clipping.** Under GRPO on MATH-500, the same definition flags 20 of 30 runs, all but one on the Qwen2.5 models (Appendix Figure 8).

Interference arises from co-optimization. The accuracy drop could come from co-optimizing the objectives or simply from the instability of RL fine-tuning, which would arise even with a single objective. To disentangle the two, we train each objective in isolation on MATH-500 with Qwen2.5-1.5B-Base (Table 2). Trained alone, accuracy ends at 0.535, well above the 0.359 reached under multi-objective training. Both runs reach similar early accuracy (0.51 vs. 0.47 over steps 20-40), but the accuracy-only run loses 0.059 from its peak and Linear loses 0.141, so most of the drop comes from co-optimization, not RL itself (Appendix Table 4). Training on conciseness alone drives accuracy to zero. Under clarity alone, accuracy peaks at 0.48 and ends at 0.134, with final responses near the length limit (1024 tokens).

Table 2: **Single- vs. multi-objective training on MATH-500** with Qwen2.5-1.5B-Base. *Final accuracy*: mean of the last five test evaluations.

Objectives	Final accuracy
Accuracy only	0.535 ± 0.017
Clarity only	0.134 ± 0.011
Conciseness only	0.000 ± 0.000
All three (Linear)	0.359 ± 0.050

Covariance flags some, but not all, declining objectives. Table 1 tests whether frequent negative covariance predicts a decline at test time. On IFEval, helpfulness has by far the largest negative fraction (0.42) and is the only objective besides conciseness to decline. On MATH-500, negative fractions stay below 0.01, and accuracy and clarity rise sharply. The pattern is nearly identical under all three methods, COVER included, so it does not depend on the weighting scheme. That said, the covariance sign does not explain every decline. MATH-500 accuracy falls from its peak (0.50 to 0.36 under Linear, Appendix Table 4) and XSum coverage collapses (Appendix Table 8), although almost none of their training prompts have negative covariance. We evaluate COVER thoroughly in §5.1.

5 COVER: COVariance-floor Enforced Reweighting

Motivated by the theoretical analysis of §4, we propose *COVariance-floor Enforced Reweighting (COVER)*, a plug-and-play, one-sided covariance controller. It raises an objective's scalarization weight only when the covariance between its reward and the clipped advantage weight (defined below) falls below a target, and otherwise leaves the weights unchanged (Appendix Algorithm 1).

Following Lu et al. (2026), we use weighted sum as the scalarization function in our experiments, $s_\lambda(\mathbf{x}, \mathbf{y}) \coloneqq \sum_{m=1}^{M} \lambda_m \, r_m(\mathbf{x}, \mathbf{y})$. Let $w_{k,l}^{\text{clip}}(\theta)$ denote the tokenwise clipped advantage weight (defined in Corollary C.4). We compute it under the updated policy θ on the sampled K completions. We aggregate $w_{k,l}^{\text{clip}}(\theta)$ over tokens to obtain a completion-level weight $w\left(\mathbf{x}, \mathbf{y}^{(k)}; \theta\right) \coloneqq \frac{1}{L_{\text{out}}} \sum_{l=1}^{L_{\text{out}}} w_{k,l}^{\text{clip}}(\theta)$. For each prompt $\mathbf{x}$ and objective m, COVER first computes the within-prompt empirical covariance $\widehat{\text{Cov}}_m(\mathbf{x}) \coloneqq \text{Cov}_{k=1\cdots K}\Big(r_m(\mathbf{x}, \mathbf{y}^{(k)}), w(\mathbf{x}, \mathbf{y}^{(k)}; \theta)\Big)$, then averages across prompts in the batch $c_m \coloneqq \mathbb{E}_{\mathbf{x} \text{ in batch}}\big[\widehat{\text{Cov}}_m(\mathbf{x})\big]$.

In practice, COVER treats the covariance as a diagnostic of whether the induced update direction benefits objective m, and uses it to adjust λ_m for the next policy update. To enforce a covariance safety margin, it runs an exponential moving average (EMA) of this signal and increases the scalarization weight when the covariance falls below a predefined target. Specifically, we maintain an EMA of the batch covariance, $\bar{c}_m \leftarrow (1-\tau)\bar{c}_m + \tau c_m$ and define a nonnegative deficit $\delta_m \coloneqq \big[c_m^* - \bar{c}_m\big]_+$. To keep $\lambda_m > 0$ and make the updates multiplicative, COVER sets $\lambda_m = \exp(u_m)$ and updates $u_m \leftarrow u_m + \eta_\lambda \, \delta_m$ with learning rate η_λ.

COVER acts as a one-sided covariance controller. Because $\delta_m \geq 0$, the log-weight u_m never decreases, so COVER raises λ_m only while $\bar{c}_m$ is below c_m^* and never lowers a weight. Two properties follow directly. **(1) Harmless by default.** While $\bar{c}_m \geq c_m^*$ for all m, the deficits vanish

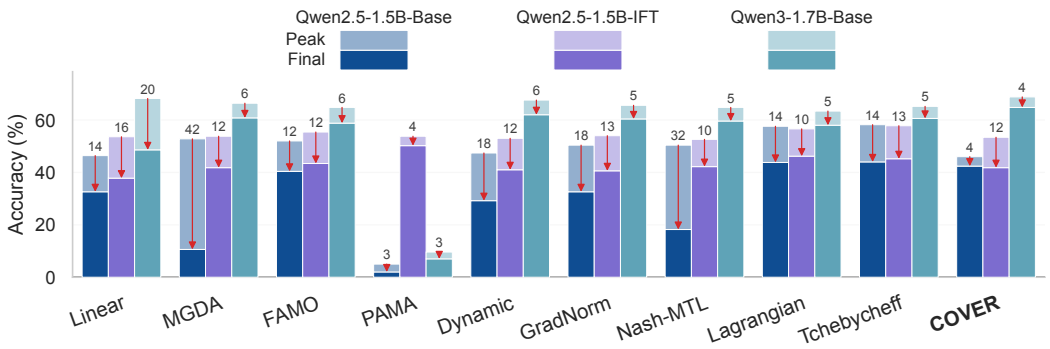


Figure 2: **Test accuracy for ten methods across three models on MATH-500.** Final accuracy averages the last five evaluations, and ↓ marks its drop from the peak. **Across models, COVER matches the strongest baselines in final accuracy with one of the smallest drops from peak.**

and the weights stop changing, so COVER acts as linear scalarization with its current weights. **(2) Early activation.** A positive target lets COVER raise a weight before the covariance turns negative.

5.1 Experiments: Evaluating COVER

COVER reduces the accuracy drop on all three models. On both Qwen2.5 models, every baseline except PAMA, whose accuracy barely moves, loses at least 0.10 from peak accuracy, up to 0.42 for MGDA (Figure 2). COVER loses 0.04 on Qwen2.5-1.5B-Base and 0.12 on Qwen2.5-1.5B-IFT, against 0.14 and 0.16 for Linear. On the Base model, only Tchebycheff and Lagrangian end higher (0.44 vs. 0.42), but their responses are three to four times longer (Figure 1a). On Qwen3-1.7B-Base, where only Linear drops more than 0.06, COVER reaches the highest final accuracy (0.65).

No harm when objectives co-improve. When objectives improve together, COVER matches Linear even though its controller stays active. On GSM8K, accuracy and clarity rise under Linear with almost no negative covariance (Appendix Table 8). COVER raises the accuracy weight to 44% of the total, while accuracy ends near Linear (0.830 vs. 0.848), clarity is 1.000 for both, and conciseness is higher (0.932 vs. 0.884). The same holds under GRPO on Qwen3-1.7B-Base, where only one of ten methods interferes and COVER matches Linear at 0.62 (Appendix Figure 8). Reweighting co-improving objectives leaves each at most 0.02 below Linear.

Higher targets reduce the accuracy drop. We raise c^*_{acc} from the usual range of $\bar{c}_{\text{acc}}$ (0.05-0.16) to well above it (Table 3 and Figure 3). Targets of 0.20 and above are never reached, so the controller raises the accuracy weight on every step. At $c^*_{\text{acc}} = 0.40$ the weight reaches 74% of the total, the drop falls to 0.101, and final accuracy reaches 0.435, though at the cost of conciseness. Peak accuracy stays between 0.50 and 0.54 across targets, so most of the gain in final accuracy comes from a smaller drop (Appendix Table 5).

Table 3: **Effect of COVER's covariance target on accuracy** on MATH-500 with Qwen2.5-1.5B-Base. *Final accuracy*: mean of the last five evaluations. *Drop*: peak minus final accuracy.

Target c^*_{acc}	Final accuracy ↑	Drop ↓
0.00 (Linear)	0.358 ±0.033	0.149 ±0.032
0.15	0.351 ±0.032	0.165 ±0.053
0.20	0.389 ±0.017	0.128 ±0.017
0.30	0.383 ±0.027	0.117 ±0.041
0.40	0.435 ±0.042	0.101 ±0.018

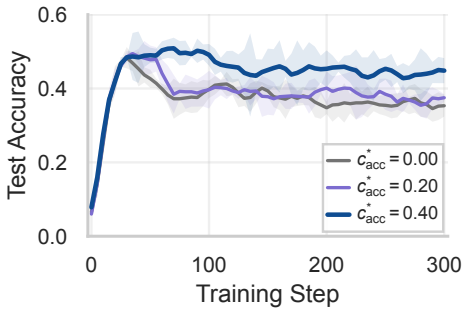


Figure 3: **Test accuracy for three targets.**

Beyond MATH-500. COVER protects an objective that declines slowly enough for the controller to respond and otherwise matches Linear. On MBPP (Qwen2.5-1.5B-IFT), COVER cuts the pass-rate drop from 0.141 to 0.090 at the cost of conciseness. On IFEval, where most baselines and fixed non-uniform weights trade one objective for another, COVER, like Linear and Dynamic, keeps all

three objectives above 0.69 (Appendix Tables 6 and 7). On Qwen2.5-7B-Base, only the 0.10 target both roughly halves the drop and ends above Linear (Appendix Table 9). Unlike Dynamic weighting or MGDA, COVER requires no per-objective gradients or projected gradient descent. The covariance statistics are computed alongside the standard RFT update with negligible overhead. However, COVER does have limits. It neither catches nor reverses the XSum coverage collapse, and at 7B only a mild covariance target avoids the accuracy collapse that also affects Linear (Appendix Tables 8 and 9).

6 Global Convergence of Multi-Objective Alignment via μ-PL Condition

While §4 characterizes when individual objectives improve locally, it does not explain the model dependence observed in practice. We therefore study the global geometry of the scalarization-induced value $V(\mathbf{x};\theta)$. Since V is highly non-convex, we ask *when the scalarized RFT objective satisfies the Polyak–Łojasiewicz (PL) condition*, a benign non-convex structure that still yields convergence guarantees (Mei et al., 2020). The resulting PL constant μ also suggests a second interference mechanism: when the geometry is unfavorable (small μ), optimization can stall near suboptimal solutions that favor easy objectives.

The analysis rests on three assumptions (formal statements in Appendix D.7): bounded scores, non-saturated token probabilities, and positively aligned token-level gradients. The first two are mild in our setting (rewards in $[0,1]$, KL regularization prevents saturation). The third is stronger and model-dependent: it requires that token-level gradient contributions along a favorable trajectory reinforce rather than cancel each other. Together with the conditioning of the logit Jacobian, this alignment is how model geometry enters μ.

Theorem 6.1 (μ-PL condition of multi-objective alignment, simplified)**.** *Under Theorems D.3 to D.5, for every prompt* $\mathbf{x}$ *the scalarized value satisfies* $\frac{1}{2}\|\nabla_\theta V(\mathbf{x};\theta)\|^2 \geq \mu\left(V(\mathbf{x};\theta^*) - V(\mathbf{x};\theta)\right)$ *with* $\mu = \frac{1}{2B}\left(\frac{p_\theta(\mathbf{y}^*|\mathbf{x})}{1-p_\theta(\mathbf{y}^*|\mathbf{x})}\, s(\mathbf{x},\mathbf{y}^*)\,\gamma - 2B\sigma_{\max}\right)$*, where* γ *collects the alignment, non-saturation, and Jacobian-conditioning constants and* $\sigma_{\max}$ *is the largest singular value of the logit Jacobian (formal statement and proof: Theorem D.6, Appendix D.7).*

The constant μ quantifies how sharply the scalarized landscape rises around its maximizer. It is large when (i) the policy places large mass on the optimal completion, so that $\frac{p_\theta(\mathbf{y}^*|\mathbf{x})}{1-p_\theta(\mathbf{y}^*|\mathbf{x})}$ is large, (ii) the optimal scalarized score $s(\mathbf{x},\mathbf{y}^*)$ is large, and (iii) the logit map is well-conditioned (large γ, which summarizes the Jacobian and non-saturation constants). Intuitively, μ reflects the net effect of the aligned ascent signal from the optimal completion minus the worst-case destructive contribution from all other completions. It therefore turns non-positive when the aligned signal is weak (small $p_\theta(\mathbf{y}^* \mid \mathbf{x})$ or small γ) or when the logit Jacobian is highly skewed (large $\sigma_{\max}/\sigma_{\min}$), so the same scalarized update can converge on one model and stall on another.

Importantly, a favorable μ ensures convergence of V but does not by itself prevent cross-objective interference: V can rise while an individual objective r_m falls whenever the covariance condition from §4 fails for that objective. Taken together, §4 and §6 disentangle two complementary mechanisms and suggest a principled recipe: *ensure convergence of the scalarized objective (make μ positive and sufficiently large) while maintaining per-objective covariance alignment along training*, so that increasing V translates into simultaneous improvement across all objectives rather than interference.

7 Conclusion

In this paper, we conducted, to our knowledge, the first systematic study of scalarization algorithms in multi-objective LLM alignment and formalized a common failure mode, cross-objective interference. Through rigorous analysis, we identified the conditions under which each objective can be improved at first order and when the scalarized optimization satisfies the PL inequality, and together these results uncover the fundamental challenges in multi-objective alignment. Guided by these insights, we proposed COVER, a lightweight covariance controller that effectively mitigates cross-objective interference. Limitations and future work are discussed in Appendix B.

AI USE STATEMENT

We did not use generative AI tools for research ideation, theoretical development, or methodology design. We used generative AI tools only to polish the wording of the paper for grammar and readability, and to generate visualization code for figures. We have reviewed all AI-polished text and verified that it preserves our intended meaning and claims. We take responsibility for the final content of this work, including text, claims, or figures produced with the aid of generative AI.

REFERENCES

Loubna Ben allal, Anton Lozhkov, Elie Bakouch, Gabriel Martin Blazquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Agustín Piqueres Lajarín, Hynek Kydlíček, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan Son NGUYEN, Ben Burtenshaw, Clémentine Fourrier, Haojun Zhao, Hugo Larcher, Mathieu Morlon, Cyril Zakka, Colin Raffel, Leandro Von Werra, and Thomas Wolf. SmolLM2: When smol goes big — data-centric training of a fully open small language model. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=3JiCl2A14H.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021. URL https://arxiv.org/abs/2108.07732.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

Leon Barrett and Srini Narayanan. Learning all optimal policies with multiple criteria. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 41–47, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390162. URL https://doi.org/10.1145/1390156.1390162.

Toygun Basaklar, Suat Gumussoy, and Umit Ogras. PD-MORL: Preference-driven multi-objective reinforcement learning algorithm. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zS9sRyaPFlJ.

V. Joseph Bowman. On the relationship of the tchebycheff norm and the efficient frontier of multiple-criteria objectives. In Hervé Thiriez and Stanley Zionts (eds.), *Multiple Criteria Decision Making*, pp. 76–86, Berlin, Heidelberg, 1976. Springer Berlin Heidelberg. ISBN 978-3-642-87563-2. URL https://link.springer.com/chapter/10.1007/978-3-642-87563-2_5.

Lisha Chen, Heshan Fernando, Yiming Ying, and Tianyi Chen. Three-way trade-off in multi-objective learning: Optimization, generalization and conflict-avoidance. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 70045–70093. Curran Associates, Inc., 2023. URL https://arxiv.org/abs/2305.20057.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/chen18a.html.

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. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Christoph Dann, Yishay Mansour, and Mehryar Mohri. Reinforcement learning can be more efficient with multiple rewards. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6948–6967. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/dann23a.html.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. Deepseek llm: Scaling open-source language models with longtermism, 2024. URL https://arxiv.org/abs/2401.02954.

Jean-Antoine Désidéri. Multiple-Gradient Descent Algorithm (MGDA). Research Report RR-6953, INRIA, June 2009. URL https://inria.hal.science/inria-00389811.

Yonathan Efroni, Ben Kretzu, Daniel R. Jiang, Jalaj Bhandari, Zheqing Zhu, and Karen Ullrich. Aligned multi objective optimization. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14989–15017. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/efroni25a.html.

Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pp. 109–117, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138881. doi: 10.1145/1014052.1014067. URL https://doi.org/10.1145/1014052.1014067.

Diego Feijer and Fernando Paganini. Stability of primal–dual gradient dynamics and applications to network optimization. *Automatica*, 46(12):1974–1981, 2010. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2010.08.011. URL https://www.sciencedirect.com/science/article/pii/S0005109810003596.

Heshan Devaka Fernando, Han Shen, Miao Liu, Subhajit Chaudhury, Keerthiram Murugesan, and Tianyi Chen. Mitigating gradient bias in multi-objective learning: A provably convergent approach. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=dLAYGdKTi2.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel

Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang,

Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Shangding Gu, Bilgehan Sel, Yuhao Ding, Lu Wang, Qingwei Lin, Alois Knoll, and Ming Jin. Safe and balanced: A framework for constrained multi-objective reinforcement learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(5):3322–3331, May 2025. ISSN 0162-8828. doi: 10.1109/TPAMI.2025.3528944. URL https://doi.org/10.1109/TPAMI.2025.3528944.

Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Zexu Sun, Bowen Sun, Huimin Chen, Ruobing Xie, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Controllable preference optimization: Toward controllable multi-objective alignment. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1437–1454, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.85. URL https://aclanthology.org/2024.emnlp-main.85/.

Qiang He and Setareh Maghsudi. Pareto multi-objective alignment for language models. In Rita P. Ribeiro, Bernhard Pfahringer, Nathalie Japkowicz, Pedro Larrañaga, Alípio M. Jorge, Carlos Soares, Pedro H. Abreu, and João Gama (eds.), *Machine Learning and Knowledge Discovery in Databases. Research Track*, pp. 257–272, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-06078-5. URL https://arxiv.org/abs/2508.07768.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization, 2025. URL https://arxiv.org/abs/2501.03262.

Zeou Hu and Yaoliang Yu. Leveraging variable sparsity to refine pareto stationarity in multi-objective optimization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Bl3e8HV9xW.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Juntao Dai, Boren Zheng, Tianyi Qiu, Jiayi Zhou, Kaile Wang, Boxuan Li, Sirui Han, Yike Guo, and Yaodong Yang. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference, 2025. URL https://arxiv.org/abs/2406.15513.

Yuxian Jiang, Yafu Li, Guanxu Chen, Dongrui Liu, Yu Cheng, and Jing Shao. Rethinking entropy regularization in large reasoning models, 2025. URL https://arxiv.org/abs/2509.25133.

Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Sai Surya Duvvuri, Manzil Zaheer, Inderjit S. Dhillon, David Brandfonbrener, and Rishabh Agarwal. The art of scaling reinforcement learning compute for llms, 2025. URL https://arxiv.org/abs/2510.13786.

Dohyeong Kim, Mineui Hong, Jeongho Park, and Songhwai Oh. Conflict-averse gradient aggregation for constrained multi-objective reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ogXkmugNZw.

Kimi, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Weixin Xu, Wenhao Wu, Wenyang He, Xianghui Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, Zonghan Yang, and Zongyu Lin. Kimi k1.5: Scaling reinforcement learning with llms, 2025. URL https://arxiv.org/abs/2501.12599.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025. URL https://arxiv.org/abs/2411.15124.

Chengao Li, Hanyu Zhang, Yunkun Xu, Hongyan Xue, Xiang Ao, and Qing He. Gradient-adaptive policy optimization: Towards multi-objective alignment of large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11214–11232, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.549. URL https://aclanthology.org/2025.acl-long.549/.

Zhuo Li, Guodong DU, Weiyang Guo, Yigeng Zhou, Xiucheng Li, Wenya Wang, Fangming Liu, Yequan Wang, Deheng Ye, Min Zhang, and Jing Li. Multi-objective large language model alignment with hierarchical experts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=UhmEdfAk46.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Baijiong Lin, Weisen Jiang, Feiyang Ye, Yu Zhang, Pengguang Chen, Ying-Cong Chen, Shu Liu, Ivor W. Tsang, and James T. Kwok. Dual-balancing for multi-task learning, 2025. URL https://arxiv.org/abs/2308.12029.

Xi Lin, Xiaoyuan Zhang, Zhiyuan Yang, Fei Liu, Zhenkun Wang, and Qingfu Zhang. Smooth tchebycheff scalarization for multi-objective optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 18878–18890. Curran Associates, Inc., 2021. URL https://arxiv.org/abs/2110.14048.

Bo Liu, Yihao Feng, Peter Stone, and qiang liu. FAMO: Fast adaptive multitask optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=zMeemcUeXL.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. GDPO: Group reward-decoupled normalization policy optimization for multi-reward RL optimization. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Y1rHxA58If.

Suyun Liu and Luis Nunes Vicente. The stochastic multi-gradient algorithm for multi-objective optimization and its application to supervised machine learning, 2021. URL https://arxiv.org/abs/1907.04472.

Haoye Lu, Daniel Herman, and Yaoliang Yu. Multi-objective reinforcement learning: Convexity, stationarity and pareto optimality. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TjEzIsyEsQ6.

Yining Lu, Zilong Wang, Shiyang Li, Xin Liu, Changlong Yu, Qingyu Yin, Zhan Shi, Zixuan Zhang, and Meng Jiang. Learning to optimize multi-objective alignment through dynamic reward weighting, 2026. URL https://arxiv.org/abs/2509.11452.

Mehrdad Mahdavi, Tianbao Yang, and Rong Jin. Stochastic convex optimization with multiple objectives. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/3493894fa4ea036cfc6433c3e2ee63b0-Paper.pdf.

Gaurav Maheshwari and Michaël Perrot. Fairgrad: Fairness aware gradient descent. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=0f8tU3QwWD.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6820–6829. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/mei20b.html.

Mistral AI Team. Ministral-8b-instruct-2410. https://huggingface.co/mistralai/Ministral-8B-Instruct-2410, 2024.

Ted Moskovitz, Aaditya K Singh, DJ Strouse, Tuomas Sandholm, Ruslan Salakhutdinov, Anca Dragan, and Stephen Marcus McAleer. Confronting reward model overoptimization with constrained RLHF. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gkfUvn0fLU.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1206. URL https://aclanthology.org/D18-1206/.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16428–16446. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/navon22a.html.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=aExAsh1UHZo.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Noam Razin, Hattie Zhou, Omid Saremi, Vimal Thilak, Arwen Bradley, Preetum Nakkiran, Joshua M. Susskind, and Etai Littwin. Vanishing gradients in reinforcement finetuning of language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=IcVNBR7qZi.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Neural Information Processing Systems*, 2018. URL https://api.semanticscholar.org/CorpusID:52957972.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Yiran Shen, Yu Xia, Jonathan Chang, and Prithviraj Ammanabrolu. Simultaneous multi-objective alignment across verifiable and non-verifiable rewards, 2025. URL https://arxiv.org/abs/2510.01167.

Guangyuan SHI, Qimai Li, Wenlong Zhang, Jiaxin Chen, and Xiao-Ming Wu. Recon: Reducing conflicting gradients from the root for multi-task learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ivwZO-HnzG_.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in RLHF. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=G8LaO1P0xv.

Ralph E Steuer and Eng-Ung Choo. An interactive weighted tchebycheff procedure for multiple objective programming. *Mathematical programming*, 26(3):326–344, 1983. URL https://link.springer.com/article/10.1007/BF02591870.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. Arithmetic control of LLMs for diverse user preferences: Directional preference alignment with multi-objective rewards. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.),

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 8642–8655, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.468. URL https://aclanthology.org/2024.acl-long.468/.

Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=F1vEjWK-lH_.

Susan Wei and Marc Niethammer. The fairness-accuracy pareto front, 2021. URL https://arxiv.org/abs/2008.10797.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc. URL https://openreview.net/forum?id=CSbGXyCswu.

Peiyao Xiao, Hao Ban, and Kaiyi Ji. Direction-oriented multi-objective learning: Simple and provable stochastic algorithms. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 4509–4533. Curran Associates, Inc., 2023. URL https://arxiv.org/abs/2305.18409.

Guofu Xie, Xiao Zhang, Ting Yao, and Yunsheng Shi. Bone soups: A seek-and-soup model merging approach for controllable multi-objective generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27237–27263, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1322. URL https://aclanthology.org/2025.acl-long.1322/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Feng Yao, Zilong Wang, Liyuan Liu, Junxia Cui, Li Zhong, Xiaohan Fu, Haohui Mai, Vish Krishnan, Jianfeng Gao, and Jingbo Shang. Training language models to generate quality code with program analysis feedback. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=3nza35A6I4.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/3fe78a8acf5fda99de95303940a2420c-Paper.pdf.

Jixiao Zhang and Chunsheng Zuo. GRPO-LEAD: A difficulty-aware reinforcement learning approach for concise mathematical reasoning in language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5642–5654, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.287. URL https://aclanthology.org/2025.emnlp-main.287/.

Xiaoyuan Zhang, Xi Lin, and Qingfu Zhang. Pmgda: A preference-based multiple gradient descent algorithm, 2024. URL https://arxiv.org/abs/2402.09492.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

APPENDIX

A Additional Experimental Results **19**
A.1 Objectives 19
A.2 Cross-Objective Interference 20
A.3 COVER Evaluation 22
A.4 Generalization and Robustness 24

B Limitations and Future Work **27**

C Formal Analysis for Clipped Surrogate Objectives **27**

D Proof **29**
D.1 Proof of Lemma D.1 29
D.2 Proof of Theorem 4.1 30
D.3 Proof of Lemma C.1 30
D.4 Proof of Theorem C.2 31
D.5 Proof of Corollary C.3 31
D.6 Proof of Corollary C.4 33
D.7 Proof of Theorem D.6 33

E Algorithms **38**

F Hyperparameters **41**

A ADDITIONAL EXPERIMENTAL RESULTS

A.1 OBJECTIVES

Every reward except harmlessness is a rule-based function of the response. Harmlessness uses a fixed toxicity classifier, and no reward relies on an LLM judge or a learned preference model, to prevent prediction noise.

MATH-500. All three rewards are binary and follow Lu et al. (2026). *Accuracy* checks whether the final `\boxed{...}` answer matches the reference. *Clarity* applies a regular-expression test for step-by-step structure. *Conciseness* returns 1 if the response is shorter, in tokens, than the running average length of all training rollouts so far, and 0 otherwise. Before the first training step no running average exists, so every response scores 1 and conciseness starts at its ceiling of 1.0.

IFEval. *Constraint following* is the fraction of verifiable instructions the response satisfies, using the IFEval checkers. *Helpfulness* is a saturating word count, $\min(\text{words}/150,\ 1)$, that serves as a rule-based proxy for response quality. We adopt length as the signal because Singhal et al. (2024) show that most reward gains in helpfulness-oriented RLHF come from longer responses, and that a reward based on length alone recovers most of the improvement of RLHF over supervised fine-tuning. We cap the reward at 150 words so that it credits a developed answer but does not reward verbosity. *Conciseness* uses the same definition as on MATH-500. Helpfulness and conciseness therefore pull response length in opposite directions.

Others. **GSM8K** uses the three MATH-500 objectives (accuracy, clarity, conciseness) on grade-school word problems. **MBPP** pairs *pass rate* (1 if the generated program passes all test assertions in a sandbox, 0 otherwise) with conciseness. **PKU-SafeRLHF** combines *harmlessness*, defined as $1 - P(\text{harmful})$ from an off-the-shelf toxicity classifier (`unitary/toxic-bert`), a *helpfulness* proxy $\min(\text{words}/120, 1)$, and conciseness. **XSum** combines *coverage* (ROUGE-L recall against the reference summary), *faithfulness* (ROUGE-L precision, a lexical proxy), and conciseness. Unlike the other domains, where we deliberately pair conflicting objectives, XSum presents a natural conflict because recall favors adding content while precision penalizes unsupported content.

A.2 Cross-Objective Interference

Definition A.1 (Operational definition of cross-objective interference)**.** Objective m suffers *cross-objective interference* if (i) m degrades from its peak by more than both a noise floor, defined as the median absolute deviation of its one-step reward differences, and 15% of its observed reward range, and (ii) at least one other objective improves after m peaks. Condition (i) alone indicates reward degradation but not interference, and an objective that merely plateaus fails (i).

We set the range threshold at 15% because it falls in a clear gap between accuracy-only and multi-objective training in Table 4. Accuracy-only training loses at most 11% of its accuracy range from peak, while every Linear, Dynamic, and COVER run in that table loses 23-45%.

Table 4: **Single- vs. multi-objective training on MATH-500** with Qwen2.5-1.5B-Base using REINFORCE. *Initialization*: the base model before training. *Early*: mean test accuracy over steps 20-40. *Final*: mean of the last five test evaluations. *Drop*: peak minus final accuracy. *Length*: mean response length in tokens. Accuracy-only and multi-objective training reach similar early accuracy (0.51 vs. 0.46-0.47), but accuracy-only training then loses only 0.059 from its peak, less than half the 0.135-0.157 lost under Linear, Dynamic, and COVER, so most of the drop comes from co-optimization.

	Accuracy			Clarity	Conciseness	Length
Objectives	Early	Final ↑	Drop ↓	Final	Final	Final
Initialization		0.020		0.030	1.000	989
Accuracy only	0.506 ±0.016	0.535 ±0.017	0.059 ±0.014	0.893 ±0.026	0.597 ±0.019	505 ±8
Clarity only	0.416 ±0.038	0.134 ±0.011	0.349 ±0.015	0.948 ±0.026	0.007 ±0.006	1020 ±5
Conciseness only	0.003 ±0.003	0.000 ±0.000	0.153 ±0.060	0.000 ±0.000	1.000 ±0.000	2 ±1
All three (Linear)	0.465 ±0.013	0.359 ±0.050	0.141 ±0.037	0.992 ±0.007	0.969 ±0.008	113 ±15
All three (Dynamic)	0.457 ±0.010	0.361 ±0.018	0.135 ±0.024	0.994 ±0.009	0.949 ±0.009	112 ±6
All three (COVER)	0.466 ±0.009	0.353 ±0.019	0.157 ±0.019	0.996 ±0.004	0.945 ±0.024	120 ±6

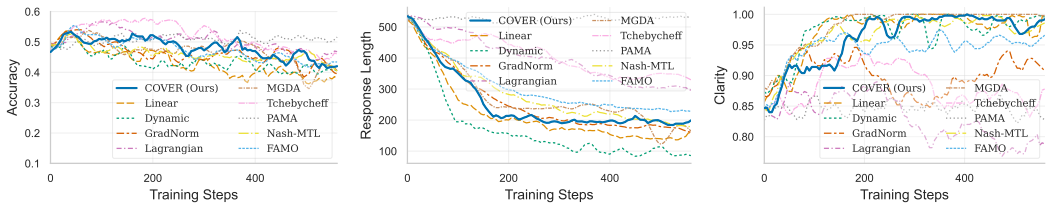


Figure 4: **MATH-500 test performance over REINFORCE training on Qwen2.5-1.5B-IFT.** All ten methods interfere with accuracy.

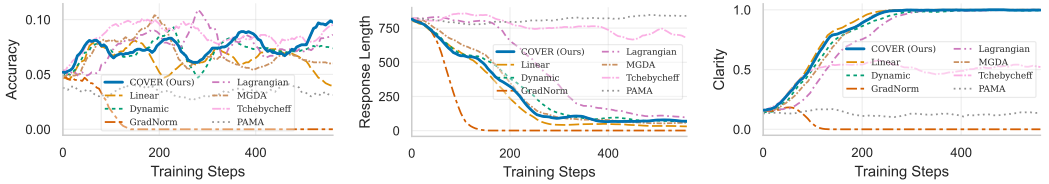
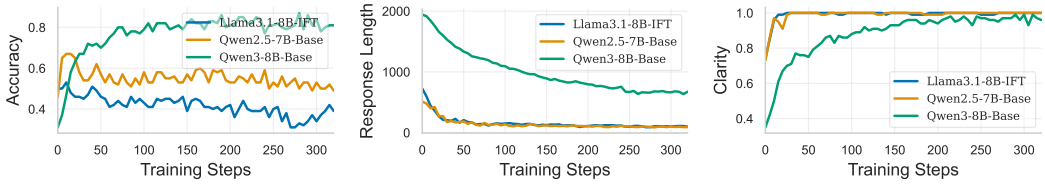
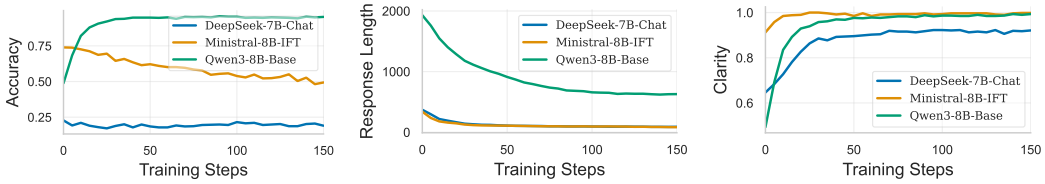


Figure 5: **MATH-500 test performance over REINFORCE training for SmolLM2-1.7B-Base** (allal et al., 2025).

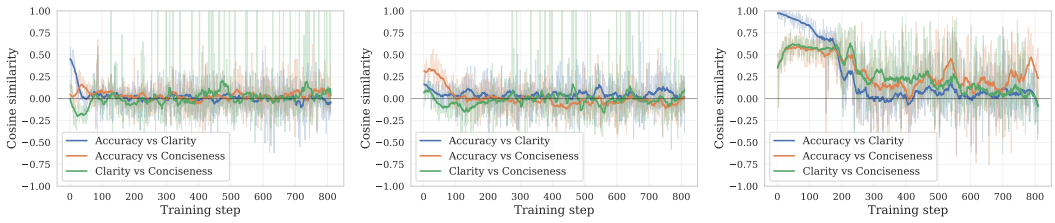


(a) Dynamic reward weighting (Lu et al., 2026) on MATH-500 (Lightman et al., 2024).



(b) Linear weighting on the MATH algebra problems (Hendrycks et al., 2021).

Figure 6: **Test performance over training at 7B-8B scale.** Columns show accuracy, response length, and clarity for five models, two datasets, and two scalarization methods. As responses shorten and clarity rises, accuracy falls from its early peak on Qwen2.5-7B-Base and ends below its starting value on Llama3.1-8B-IFT, Ministral-8B-IFT, and DeepSeek-7B-Chat, whereas Qwen3-8B-Base co-improves on both datasets. **Interference is not a small-model artifact.**



(a) Qwen2.5-1.5B-Base (b) Qwen2.5-1.5B-IFT (c) Qwen3-1.7B-Base

Figure 7: **Gradient alignment across objectives during multi-objective alignment.** We measure pairwise cosine similarity between per-objective gradients throughout training. Negative values indicate conflicting updates, which is a standard proxy for identifying conflicting objectives in MTL (Yu et al., 2020). Across all three models, cosine similarities remain mostly non-negative and converge toward 0 as training progresses, suggesting that objectives are weakly coupled, neither strongly synergistic nor persistently antagonistic, with no conflicting behavior observed.

A.3 COVER EVALUATION

Table 5: **Effect of COVER's covariance target on accuracy** on MATH-500 using Qwen2.5-1.5B-Base, extending Table 3 with additional targets. λ_{acc}: unnormalized accuracy weight. *Final accuracy* and *Drop* are defined as in Table 3. The drop falls below that of Linear only at targets of 0.20 and above, where the controller raises the accuracy weight the most.

Target c^*_{acc}	λ_{acc} (start $\rightarrow$ end)	Peak accuracy	Final accuracy ↑	Drop ↓
0.00 (Linear)	$0.333 \rightarrow 0.333$	0.507 ± 0.018	0.358 ± 0.033	0.149 ± 0.032
0.05	$0.333 \rightarrow 0.334$	0.520 ± 0.010	0.369 ± 0.029	0.151 ± 0.028
0.10	$0.333 \rightarrow 0.351$	0.510 ± 0.044	0.352 ± 0.018	0.158 ± 0.042
0.15	$0.333 \rightarrow 0.408$	0.517 ± 0.021	0.351 ± 0.032	0.165 ± 0.053
0.20	$0.333 \rightarrow 0.573$	0.517 ± 0.012	0.389 ± 0.017	0.128 ± 0.017
0.30	$0.333 \rightarrow 1.072$	0.500 ± 0.020	0.383 ± 0.027	0.117 ± 0.041
0.40	$0.333 \rightarrow 1.925$	0.537 ± 0.025	0.435 ± 0.042	0.101 ± 0.018

Table 6: **Multi-objective IFEval results** on Qwen2.5-1.5B-IFT (mean of the last five evaluations). Helpfulness rewards length up to 150 words, while conciseness rewards responses shorter than the running mean length of the method's own training rollouts (Appendix A.1), so the two conflict only once that mean falls below the length at which helpfulness saturates. Lagrangian reaches the highest constraint following but violates its own 0.85 conciseness constraint. MGDA and FAMO saturate helpfulness at the cost of constraint following. Tchebycheff collapses to one-token responses and PAMA to maximum-length responses, so each fails on two objectives. Only Linear, Dynamic, and COVER keep all three objectives above 0.69, and these three methods differ by at most 0.03 on each objective.

Method	Helpfulness ↑	Constraint following ↑	Conciseness ↑
Linear	0.715 ± 0.032	0.693 ± 0.039	0.904 ± 0.027
Dynamic	0.718 ± 0.021	0.707 ± 0.027	0.899 ± 0.025
Lagrangian	0.736 ± 0.028	0.755 ± 0.026	0.582 ± 0.033
Tchebycheff	0.000 ± 0.000	0.241 ± 0.000	1.000 ± 0.000
PAMA	0.969 ± 0.040	0.258 ± 0.064	0.068 ± 0.094
MGDA	1.000 ± 0.000	0.541 ± 0.026	0.959 ± 0.019
GradNorm	0.921 ± 0.123	0.675 ± 0.088	0.772 ± 0.193
Nash-MTL	0.628 ± 0.307	0.674 ± 0.072	0.961 ± 0.024
FAMO	0.998 ± 0.002	0.485 ± 0.016	0.987 ± 0.010
COVER	0.729 ± 0.026	0.714 ± 0.018	0.890 ± 0.024

Table 7: **Static-weight ablation on IFEval,** comparing fixed weight vectors against COVER on Qwen2.5-1.5B-IFT. The reported numbers are the mean of the last five evaluations. Weight vectors are ordered as helpfulness, constraint following, and conciseness. The vector $[.33, .34, .33]$ is the Linear baseline, and $[.34, .35, .31]$ fixes each weight at its average normalized value under COVER. Every fixed vector that departs from Linear raises helpfulness but lowers constraint following below Linear, whereas COVER reaches the highest constraint following in this table, above its own average weights held fixed (0.714 vs. 0.676), which suggests that the timing of its reweighting matters and not only its average weights.

Weights	Helpfulness ↑	Constraint following ↑	Conciseness ↑
$[.33, .34, .33]$	0.715 ± 0.032	0.693 ± 0.039	0.904 ± 0.027
$[.34, .35, .31]$	0.764 ± 0.001	0.676 ± 0.020	0.863 ± 0.041
$[.45, .30, .25]$	0.954 ± 0.011	0.626 ± 0.003	0.923 ± 0.019
$[.55, .25, .20]$	0.993 ± 0.003	0.612 ± 0.010	0.957 ± 0.013
COVER (Adaptive)	0.729 ± 0.026	0.714 ± 0.018	0.890 ± 0.024

A.4 Generalization and Robustness

Table 8: **Interference extends beyond math and instruction following,** using Qwen2.5-1.5B-IFT. *Mean covariance* (in units of 10^{-2}) and *Negative fraction* are defined as in Table 1. *Start* is the first evaluation, *Peak* the best, and *Final* the mean of the last five. Start separates objectives that never rose (MBPP pass rate, XSum coverage) from those that rose and plateaued (GSM8K accuracy, XSum faithfulness). *Engaged* is the fraction of training steps on which COVER raised the objective's weight, and *Final* λ_m is its unnormalized weight at the end of training. Weights start uniform (0.33 each, or 0.5 on MBPP). Covariance targets are $c^* = 0.15$ for the task objective and 0.05 for the rest, with $\eta_\lambda = 0.02$. †Conciseness starts at its ceiling of 1.0 (Appendix A.1).

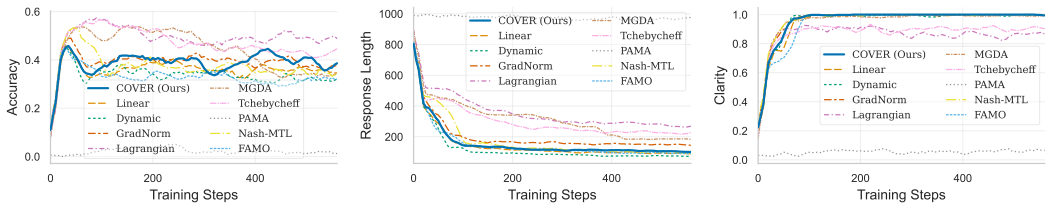
		Mean covariance	Negative fraction	Test performance						COVER	
				Linear			COVER			Engaged	Final
Dataset	Objective	Linear	Linear	Start	Peak	Final	Start	Peak	Final	(%)	λ_m
GSM8K	Accuracy	1.44	0.000	0.080	0.900	0.848	0.080	0.910	0.830	76.8	0.520
	Clarity	0.51	0.000	0.060	1.000	1.000	0.060	1.000	1.000	0.0	0.334
	Conciseness†	0.84	0.001	1.000	1.000	0.884	1.000	1.000	0.932	10.8	0.338
MBPP	Pass rate	0.81	0.000	0.565	0.565	0.424	0.565	0.565	0.475	100.0	1.005
	Conciseness†	0.85	0.000	1.000	1.000	0.864	1.000	1.000	0.720	86.1	0.594
PKU-SafeRLHF	Harmlessness	0.00	0.439	0.999	0.999	0.999	0.999	0.999	0.999	100.0	2.248
	Helpfulness	−0.01	0.286	0.390	0.722	0.717	0.390	0.749	0.742	99.7	0.609
	Conciseness†	1.11	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.2	0.330
XSum	Coverage	0.04	0.019	0.241	0.241	0.054	0.246	0.246	0.053	100.0	1.573
	Faithfulness	0.30	0.001	0.155	0.869	0.794	0.158	0.831	0.737	44.6	0.378
	Conciseness†	0.13	0.000	1.000	1.000	0.944	1.000	1.000	0.961	11.4	0.348

Covariance is large while an objective improves and small once it plateaus. Equation (3) predicts this pattern at first order, and the Linear training logs behind Table 8 show it. On GSM8K, test accuracy rises from 0.08 to 0.90 over the first 35 steps and then plateaus. Its covariance, in units of 10^{-2}, averages 5.7 during the rise and 0.91 afterward. XSum faithfulness follows the same pattern. It rises early to about 0.45, then climbs again to 0.74 between steps 90 and 115 before plateauing near 0.80. Its covariance averages 0.84 during this second climb and 0.19 afterward. The exceptions are MBPP pass rate and XSum coverage, which decline even though their covariance stays positive.

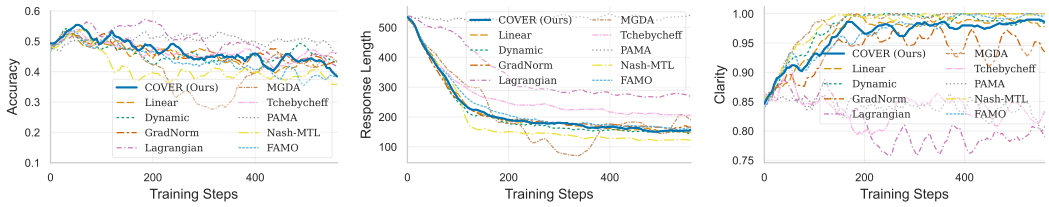
COVER slows a gradual decline but cannot reverse a collapse. On MBPP, the pass rate declines gradually from its first evaluation, and the controller raises its weight on every step, from 50% to 63% of the total by the end of training. COVER slows the decline, reducing the drop from 0.141 to 0.090 at the cost of conciseness, which ends at 0.720 compared with 0.864 under Linear. XSum coverage, in contrast, loses nearly half its value in the first ten steps and collapses again near step 100 as summaries shrink to about 4 tokens. Although the controller raises the coverage weight to 68% of the total by the end of training, coverage ends level with Linear (0.053 vs. 0.054) and faithfulness ends lower (0.737 vs. 0.794). The other two domains show no interference, so there is nothing for the controller to slow. On GSM8K, where objectives co-improve, COVER ends with accuracy close to Linear (0.830 vs. 0.848). On PKU-SafeRLHF, harmlessness is saturated at 0.999, so its covariance never reaches the target. The controller keeps raising its weight, but the near-constant reward adds little to the update.

Table 9: **Effect of COVER's covariance target on accuracy at 7B** on MATH-500 using Qwen2.5-7B-Base. *Peak*: best test accuracy up to step 210. *Accuracy*: test accuracy at step 210. *Drop*: Peak minus Accuracy. *Final*: mean test accuracy over the last five evaluations of the full run. Unlike at 1.5B (Table 5), where the highest target works best, only the 0.10 target reliably reduces the drop and stays above Linear through the full run, suggesting that larger models need more conservative targets to avoid destabilizing training.

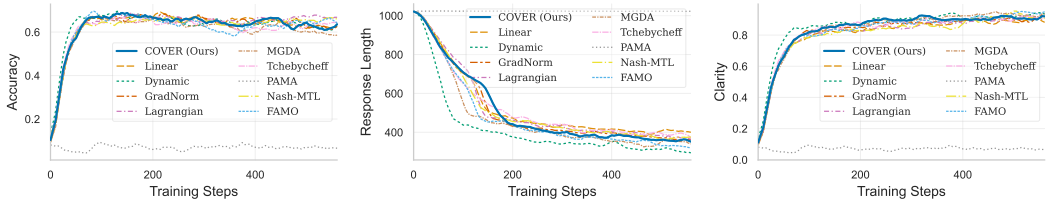
	Steps 0-210			Full run (315 steps)
Target c^*_{acc}	Peak ↑	Accuracy ↑	Drop ↓	Final ↑
0.00 (Linear)	0.663 ±0.006	0.307 ±0.171	0.357 ±0.169	0.299 ±0.212
0.10	0.643 ±0.029	0.457 ±0.025	0.187 ±0.051	0.449 ±0.049
0.20	0.657 ±0.015	0.483 ±0.035	0.173 ±0.032	0.276 ±0.247
0.30	0.653 ±0.015	0.433 ±0.178	0.220 ±0.165	0.453 ±0.118
0.40	0.653 ±0.006	0.160 ±0.277	0.493 ±0.280	0.401 ±0.126



(a) Qwen2.5-1.5B-Base



(b) Qwen2.5-1.5B-IFT



(c) Qwen3-1.7B-Base

Figure 8: **MATH-500 test performance over GRPO training for Qwen2.5 and Qwen3 models.** Applying Theorem A.1, all ten methods interfere with accuracy on Qwen2.5-1.5B-Base (PAMA only nominally, since it never learns the task), nine of ten on Qwen2.5-1.5B-IFT, and only one of ten on Qwen3-1.7B-Base. The same pattern holds under REINFORCE, so clipping does not change which models suffer interference. COVER matches Linear on every model, whereas Lagrangian and Tchebycheff keep higher accuracy but pay with longer responses and lower clarity.

Table 10: **Effect of spectral decoupling on MATH-500** with Qwen2.5-1.5B-Base using REINFORCE. Spectral decoupling (Pezeshki et al., 2021) adds the logit-norm penalty $\frac{\kappa}{2}\|\text{logits}\|^2$ to the policy loss as a remedy for gradient starvation. *Gradient norm*: median gradient norm over the first five updates, measured before gradient-norm clipping to 1. *Final*: mean of the last five test evaluations. *Final entropy*: policy entropy on the training batch at the last update, whose maximum $\ln|\mathcal{X}| = \ln 151936 = 11.93$ is that of the uniform distribution over the vocabulary. Dashes mark unavailable quantities. Because the penalty sums over the full vocabulary, even $\kappa = 10^{-4}$ raises the gradient norm to 1,072, about 2,500 times the 0.43 of the unpenalized accuracy-only run. Every penalized run reaches this maximum entropy, including the accuracy-only runs, which have no competing objective. The collapse therefore comes from the penalty itself, not from cross-objective interference. **At every tested strength, the logit-norm remedy for gradient starvation collapses the policy.**

			Test accuracy			
Objectives	Penalty κ	Gradient norm	Peak (step)	Step 25	Final ↑	Final entropy
Accuracy only	0	0.43	0.610 (165)	0.440	0.526	0.11
Accuracy only	3×10^{-4}	3,160	0.200 (5)	0.000	0.000	11.93
Accuracy only	10^{-3}	10,971	0.190 (5)	0.020	0.000	11.93
Accuracy only	3×10^{-3}	28,495	0.220 (10)	0.000	0.000	11.93
Accuracy only	10^{-2}	109,646	0.190 (10)	0.000	0.000	11.93
All three (Linear)	0	–	0.520 (50)	0.440	0.370	–
All three (Linear)	10^{-4}	1,072	0.220 (10)	0.000	0.000	11.93
All three (Linear)	3×10^{-4}	2,845	0.200 (10)	0.020	0.000	11.93
All three (Linear)	10^{-3}	9,362	0.210 (15)	0.000	0.000	11.93
All three (Linear)	3×10^{-3}	32,148	0.190 (10)	0.000	0.000	11.93
All three (Linear)	10^{-2}	107,128	0.180 (10)	0.000	0.000	11.93

B Limitations and Future Work

When COVER helps. COVER helps only when an objective declines gradually, and even then it reallocates performance rather than yielding a Pareto improvement (Appendix A.4). When objectives co-improve or one is saturated, it performs on par with Linear, and when an objective collapses, raising its weight does not bring it back. Reversing a collapse likely requires more than reweighting, and we leave it to future work.

Scope of Baselines. We select representative baselines that have either been shown to be effective for multi-objective LLM alignment, such as Dynamic Weighting, PAMA, and Linear Scalarization, or are canonical methods in MTL and MOO, such as MGDA, GradNorm, and Nash-MTL. Our goal is not to exhaustively benchmark all existing methods from the MTL, MOO, and LLM alignment literatures, but rather to uncover the limitations of major reward-level and gradient-level scalarization paradigms for LLM alignment. We leave the adaptation and evaluation of additional approaches, such as SDMGrad (Xiao et al., 2023) and FairGrad (Maheshwari & Perrot, 2023), to future work.

Practical Application. Theorem D.6 explains model-dependence of cross-objective interference, but directly turning it into a practical training strategy is challenging, because Assumption D.5 has token-level Jacobian terms that are prohibitively expensive to compute for modern LMs. We also explored an approximate solution that projects per-token logit gradients into a cone around a shared reference direction during backpropagation, but it did not yield desired gains. We therefore view Theorem D.6 mainly as an explanatory framework and we hope it may inspire more scalable training methods in future work.

C Formal Analysis for Clipped Surrogate Objectives

We establish useful lemmas and theorems that analyze sufficient conditions for multi-objective improvement under clipped surrogate objectives. We use GRPO as a running example, while the same argument applies to PPO-style (Schulman et al., 2017) clipped surrogate objectives, with the only difference being the advantage estimator. All proofs are deferred to Appendix D.

Fix a prompt $\mathbf{x}$ and sample a group of K completions $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(K)} \sim p_{\theta_{\text{old}}}(\cdot \mid \mathbf{x})$. Let $s_j \coloneqq s(\mathbf{x}, \mathbf{y}^{(j)})$ and define the group normalization $A_i(\mathbf{x}) \coloneqq \frac{s_i - \bar{s}(\mathbf{x})}{\hat{\sigma}(\mathbf{x})}$. The GRPO surrogate objective is

$$J(\theta) \coloneqq \mathbb{E}\left[\sum_{k=1}^{K}\sum_{l=1}^{L_{\text{out}}} \min\Big\{\rho_{k,l}(\theta)A_k(\mathbf{x}),\ \bar{\rho}_{k,l}(\theta)A_k(\mathbf{x})\Big\}\right] - \beta\,\mathbb{E}\big[\mathrm{KL}(p_\theta \| p_{\text{ref}})\big] + \lambda\,\mathbb{E}\big[H(p_\theta)\big],$$

where $\beta, \lambda \geq 0$ control KL and entropy regularization. $\bar{\rho}_{k,l}(\theta)$ is the clipped result of importance ratio $\rho_{k,l}(\theta)$, $\bar{\rho}_{k,l}(\theta) = \mathrm{clip}\big(\rho_{k,l}(\theta), 1-\varepsilon, 1+\varepsilon\big)$. We know that the unclipped indicator selecting the active branch of the minimum is

$$\mathbf{1}_{k,l}(\theta) = \begin{cases} 1, & A_k(\mathbf{x}) \geq 0 \text{ and } \rho_{k,l}(\theta) \leq 1+\varepsilon, \\ 1, & A_k(\mathbf{x}) < 0 \text{ and } \rho_{k,l}(\theta) \geq 1-\varepsilon, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Recall $f(\mathbf{x}, \mathbf{y}_{\leq l-1}; \theta) \in \mathbb{R}^{|\mathcal{X}|}$ is the next-token logit map and $p_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{\leq l-1}) = \mathrm{softmax}(f(\mathbf{x}, \mathbf{y}_{\leq l-1}; \theta))$. Define the tokenwise logit-gradient feature $\phi_l(\mathbf{x}, \mathbf{y}; \theta)$ and the clipped advantage weight $W_{k,l}(\theta)$:

$$\phi_l(\mathbf{x}, \mathbf{y}; \theta) \coloneqq \nabla_\theta \log p_\theta\big(y_l \mid \mathbf{x}, \mathbf{y}_{\leq l-1}\big),\ W_{k,l}(\theta) \coloneqq A_k(\mathbf{x})\,\rho_{k,l}(\theta)\,\mathbf{1}_{k,l}(\theta). \tag{7}$$

Lemma C.1. *Let $r_m(\theta) \coloneqq r_m(p_\theta)$ and assume r_m has an L_m-Lipschitz gradient around θ. For any direction $d(\theta) = \nabla J(\theta)$ and update $\theta^+ = \theta + \eta d(\theta)$, we have*

$$r_m(\theta^+) - r_m(\theta) \ \geq\ \eta\,\langle \nabla r_m(\theta), d(\theta)\rangle - \frac{L_m}{2}\eta^2 \|d(\theta)\|^2,\ \forall m.$$

In particular, if $\langle \nabla r_m(\theta), d(\theta)\rangle \geq 0$ and $\eta > 0$ is sufficiently small, then $r_m(\theta^+) \geq r_m(\theta)$.

Theorem C.2 (Fisher-covariance sufficient condition for natural gradient updates)**.** *For completions $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(K)} \sim p_{\theta_{\text{old}}}(\cdot \mid \mathbf{x})$, write $\phi_{k,l}(\theta) \coloneqq \phi_l(\mathbf{x}, \mathbf{y}^{(k)}; \theta)$. Define the aggregated Fisher matrix:*

$$F(\theta) \coloneqq \mathbb{E}\left[\Big(\sum_{k=1}^{K}\sum_{l=1}^{L_{\text{out}}} \phi_{k,l}(\theta)\Big)\Big(\sum_{k=1}^{K}\sum_{l=1}^{L_{\text{out}}} \phi_{k,l}(\theta)\Big)^{\top}\right],$$

and the weighted feature mean and regularizer gradient:

$$G(\theta) \coloneqq \mathbb{E}\left[\sum_{k=1}^{K}\sum_{l=1}^{L_{\text{out}}} W_{k,l}(\theta)\,\phi_{k,l}(\theta)\right],\ R(\theta) \coloneqq \beta\,\nabla_\theta \mathbb{E}\big[\mathrm{KL}(p_\theta \| p_{\text{ref}})\big] - \lambda\,\nabla_\theta \mathbb{E}\big[H(p_\theta)\big].$$

Assume $F(\theta)$ is invertible and consider the natural gradient direction $d_{\text{nat}}(\theta) \coloneqq F(\theta)^{-1}\big(G(\theta) - R(\theta)\big),\ \theta^+ = \theta + \eta\, d_{\text{nat}}(\theta)$. If for every m,

$$\nabla r_m(\theta)^\top F(\theta)^{-1} G(\theta) \ \geq\ \nabla r_m(\theta)^\top F(\theta)^{-1} R(\theta), \tag{8}$$

and moreover either $d_{\text{nat}}(\theta) = 0$ or the inequality holds with a positive margin

$$\gamma_m(\theta) \coloneqq \nabla r_m(\theta)^\top F(\theta)^{-1}\big(G(\theta) - R(\theta)\big) > 0,\ \forall m,$$

then for sufficiently small $\eta > 0$ we have $r_m(\theta^+) \geq r_m(\theta)$ for all m and strictly if $\min_m \gamma_m(\theta) > 0$.

Corollary C.3 (Categorical bandit case)**.** *Fix a prompt $\mathbf{x}$ and suppose the policy over completions $\mathbf{y} \in \mathcal{X}^{L_{\text{out}}}$ is a categorical distribution parameterized by logits $\theta \in \mathbb{R}^{|\mathcal{X}^{L_{\text{out}}}|}$*

$$p_\theta(\mathbf{y} \mid \mathbf{x}) = \frac{\exp(\theta_{\mathbf{y}})}{\sum_{\mathbf{y}'} \exp(\theta_{\mathbf{y}'})}.$$

Let $w(\mathbf{x}, \mathbf{y})$ be an arbitrary scalar weight assigned to each completion (i.e., a per-sample quantity whose expected value $\mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}[w(\mathbf{x}, \mathbf{y})]$ we seek to maximize via natural gradient ascent). Define the categorical Fisher matrix

$$F(\theta) \coloneqq \mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[\nabla_\theta \log p_\theta(\mathbf{y} \mid \mathbf{x}) \nabla_\theta \log p_\theta(\mathbf{y} \mid \mathbf{x})^\top\big].$$

Let $d_{\text{nat}}(\theta)$ be any natural gradient direction and take the update $\theta^+ = \theta + \eta d_{\text{nat}}(\theta)$ with learning rate $\eta > 0$. Then for each objective m, we have

$$\mathbb{E}_{\mathbf{y}\sim p_{\theta^+}(\cdot|\mathbf{x})}\big[r_m(\mathbf{x}, \mathbf{y})\big] - \mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[r_m(\mathbf{x}, \mathbf{y})\big] = \eta \mathrm{Cov}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big(r_m(\mathbf{x}, \mathbf{y}), w(\mathbf{x}, \mathbf{y})\big) + O(\eta^2).$$

Specifically, $\mathrm{Cov}\big(r_m(\mathbf{x}, \mathbf{y}), w(\mathbf{x}, \mathbf{y})\big) \geq 0$ for all m is a sufficient first-order condition to ensure that no objective degrades.

Corollary C.4 (Clipping robustness)**.** *Define the unclipped and clipped weights for each token*

$$w_{k,l}^{\text{unclip}}(\theta) \coloneqq A_k(\mathbf{x})\rho_{k,l}(\theta), w_{k,l}^{\text{clip}}(\theta) \coloneqq w_{k,l}^{\text{unclip}}(\theta)\mathbf{1}_{k,l}(\theta),$$

where $\mathbf{1}_{k,l}(\theta) \in \{0, 1\}$ is defined in Equation (6) *and $\mathbf{1}_{k,l}(\theta) = 0$ exactly when the term is clipped away and contributes zero gradient. Let*

$$G^{\text{unclip}}(\theta) \coloneqq \mathbb{E}\left[\sum_{k=1}^{K}\sum_{l=1}^{L_{\text{out}}} w_{k,l}^{\text{unclip}}(\theta)\phi_{k,l}(\theta)\right],\ G^{\text{clip}}(\theta) \coloneqq \mathbb{E}\left[\sum_{k=1}^{K}\sum_{l=1}^{L_{\text{out}}} w_{k,l}^{\text{clip}}(\theta)\phi_{k,l}(\theta)\right].$$

For each objective m, define the unclipped *first-order margin*

$$\gamma_m^{\text{unclip}}(\theta) \coloneqq \nabla r_m(\theta)^\top F(\theta)^{-1}\big(G^{\text{unclip}}(\theta) - R(\theta)\big).$$

Then the clipped first-order margin satisfies

$$\begin{aligned}\nabla r_m(\theta)^\top F(\theta)^{-1}&\big(G^{\text{clip}}(\theta) - R(\theta)\big)\\ &\geq \gamma_m^{\text{unclip}}(\theta) - \big\|F(\theta)^{-1/2}\nabla r_m(\theta)\big\| \cdot \big\|F(\theta)^{-1/2}\big(G^{\text{unclip}}(\theta) - G^{\text{clip}}(\theta)\big)\big\|.\end{aligned}$$

Consequently, if $\gamma_m^{\text{unclip}}(\theta) \geq \kappa_m > 0$ and

$$\big\|F(\theta)^{-1/2}\big(G^{\text{unclip}}(\theta) - G^{\text{clip}}(\theta)\big)\big\| \leq \frac{\kappa_m}{\big\|F(\theta)^{-1/2}\nabla r_m(\theta)\big\|},$$

then $\nabla r_m(\theta)^\top F(\theta)^{-1}\big(G^{\text{clip}}(\theta) - R(\theta)\big) \geq 0$, so the sufficient condition for $r_m(\theta^+) \geq r_m(\theta)$ in Theorem C.2 still holds for objective m under clipping.

Remark C.5. Corollary C.4 shows that clipping can only affect a first-order improvement guarantee by removing a subset of weighted logit-gradient terms $w_{k,l}^{\mathrm{unclip}}(\theta)\,\phi_{k,l}(\theta)$ from the update. The Fisher-distortion $\|F(\theta)^{-1/2}(G^{\mathrm{unclip}}(\theta) - G^{\mathrm{clip}}(\theta))\|$ therefore measures, in natural gradient geometry, how much "gradient mass" is deleted by clipping. This distortion becomes large when many importance ratios fall outside the clipping window, especially on samples with large magnitude advantages. Importantly, this can be carefully controlled in practice through techniques such as learning rate scheduling or reward normalization, ensuring that clipping distortion remains small and the covariance law continues to hold.

D PROOF

D.1 PROOF OF LEMMA D.1

Lemma D.1. *The optimizer of Equation* (2) *is* $p^{+}_{\theta;\mathbf{x}}(\mathbf{y}) = \frac{p_{\theta;\mathbf{x}}(\mathbf{y})\exp\big(\eta\, s(\mathbf{x},\mathbf{y})\big)}{\mathbb{E}_{\mathbf{y}\sim p_{\theta;\mathbf{x}}}\Big[\exp\big(\eta\, s(\mathbf{x},\mathbf{y})\big)\Big]}$.

Proof. Fix $\mathbf{x}$. We optimize over distributions q on $\mathcal{X}^{L_{\mathrm{out}}}$ satisfying $q(\mathbf{y}) \geq 0$ and $\sum_{\mathbf{y}} q(\mathbf{y}) = 1$. Expanding the KL term gives

$$\mathrm{KL}(q\|p_{\theta;\mathbf{x}}) = \sum_{\mathbf{y}} q(\mathbf{y}) \log \frac{q(\mathbf{y})}{p_{\theta;\mathbf{x}}(\mathbf{y})}.$$

Thus the objective can be written as

$$\max_{q\in\Delta(\mathcal{X}^{L_{\mathrm{out}}})} \left\{ \sum_{\mathbf{y}} q(\mathbf{y}) s(\mathbf{x},\mathbf{y}) - \frac{1}{\eta} \sum_{\mathbf{y}} q(\mathbf{y}) \log \frac{q(\mathbf{y})}{p_{\theta;\mathbf{x}}(\mathbf{y})} \right\}.$$

Introduce a Lagrange multiplier $\lambda \in \mathbb{R}$ for the normalization constraint $\sum_{\mathbf{y}} q(\mathbf{y}) = 1$, and consider

$$\mathcal{L}(q,\lambda) = \sum_{\mathbf{y}} q(\mathbf{y})\, s(\mathbf{x},\mathbf{y}) - \frac{1}{\eta} \sum_{\mathbf{y}} q(\mathbf{y}) \log \frac{q(\mathbf{y})}{p_{\theta;\mathbf{x}}(\mathbf{y})} + \lambda\Big(\sum_{\mathbf{y}} q(\mathbf{y}) - 1\Big).$$

For any $\mathbf{y}$ with $q(\mathbf{y}) > 0$, the stationarity condition is

$$\frac{\partial}{\partial q(\mathbf{y})}\mathcal{L}(q,\lambda) = s(\mathbf{x},\mathbf{y}) - \frac{1}{\eta}\left(\log \frac{q(\mathbf{y})}{p_{\theta;\mathbf{x}}(\mathbf{y})} + 1\right) + \lambda = 0.$$

Rearranging yields

$$\log \frac{q(\mathbf{y})}{p_{\theta;\mathbf{x}}(\mathbf{y})} = \eta\, s(\mathbf{x},\mathbf{y}) + \eta\lambda - 1.$$

Exponentiating both sides gives us

$$q(\mathbf{y}) = p_{\theta;\mathbf{x}}(\mathbf{y}) \exp\big(\eta\, s(\mathbf{x},\mathbf{y})\big) \exp(\eta\lambda - 1).$$

Let $C := \exp(\eta\lambda - 1)$, which does not depend on $\mathbf{y}$. Enforcing normalization $\sum_{\mathbf{y}} q(\mathbf{y}) = 1$ implies

$$1 = \sum_{\mathbf{y}} q(\mathbf{y}) = C \sum_{\mathbf{y}} p_{\theta;\mathbf{x}}(\mathbf{y}) \exp\big(\eta\, s(\mathbf{x},\mathbf{y})\big),$$

so

$$C = \frac{1}{\sum_{\mathbf{y}} p_{\theta;\mathbf{x}}(\mathbf{y}) \exp\big(\eta\, s(\mathbf{x},\mathbf{y})\big)} = \frac{1}{\mathbb{E}_{\mathbf{y}\sim p_{\theta;\mathbf{x}}}\big[\exp(\eta\, s(\mathbf{x},\mathbf{y}))\big]}.$$

Substituting back, the maximizer is

$$q(\mathbf{y}) = \frac{p_{\theta;\mathbf{x}}(\mathbf{y}) \exp\big(\eta\, s(\mathbf{x},\mathbf{y})\big)}{\mathbb{E}_{\mathbf{y}\sim p_{\theta;\mathbf{x}}}\big[\exp(\eta\, s(\mathbf{x},\mathbf{y}))\big]}.$$

Identifying $q = p^{+}_{\theta;\mathbf{x}}$ yields the claim. □

Remark D.2. The update Equation (2) trades off increasing $\mathbb{E}_{\mathbf{y}\sim q}[s(\mathbf{x},\mathbf{y})]$ with staying close to $p_{\theta;\mathbf{x}}$ in $\mathrm{KL}(q\|p_{\theta;\mathbf{x}})$. Lemma D.1 shows the solution is an exponential tilting, $p^{+}_{\theta;\mathbf{x}}(\mathbf{y}) \propto p_{\theta;\mathbf{x}}(\mathbf{y}) e^{\eta s(\mathbf{x},\mathbf{y})}$, so it can only reweight completions already supported by $p_{\theta;\mathbf{x}}$. In particular, if $p_{\theta;\mathbf{x}}(\mathbf{y}) = 0$ then $p^{+}_{\theta;\mathbf{x}}(\mathbf{y}) = 0$, since otherwise $\mathrm{KL}(q\|p_{\theta;\mathbf{x}}) = +\infty$.

D.2 Proof of Theorem 4.1

Proof. Fix $\mathbf{x}$. By Lemma D.1, the optimizer of Equation (2) satisfies

$$p^{+}_{\theta;\mathbf{x}}(\mathbf{y}) = \frac{p_{\theta;\mathbf{x}}(\mathbf{y}) \exp\big(\eta\, s(\mathbf{x},\mathbf{y})\big)}{\mathbb{E}_{\mathbf{y}\sim p_{\theta;\mathbf{x}}}\Big[\exp\big(\eta\, s(\mathbf{x},\mathbf{y})\big)\Big]}$$

Let $\bar{s}(\mathbf{x}) \coloneqq \mathbb{E}_{\mathbf{y}\sim p_{\theta;\mathbf{x}}}\big[s(\mathbf{x},\mathbf{y})\big]$. Using the Taylor expansion $\exp(\eta s) = 1 + \eta s + O(\eta^2)$, we have

$$Z_\eta(\mathbf{x}) = \mathbb{E}_{\mathbf{y}\sim p_{\theta;\mathbf{x}}}\big[1 + \eta s(\mathbf{x},\mathbf{y}) + O(\eta^2)\big] = 1 + \eta\bar{s}(\mathbf{x}) + O(\eta^2).$$

For small η, $(1+u)^{-1} = 1 - u + O(u^2)$ implies

$$\frac{1}{Z_\eta(\mathbf{x})} = 1 - \eta\,\bar{s}(\mathbf{x}) + O(\eta^2).$$

Substituting back yields the pointwise expansion

$$p^{+}_{\theta;\mathbf{x}}(\mathbf{y}) = p_{\theta;\mathbf{x}}(\mathbf{y})\Big(1 + \eta\big(s(\mathbf{x},\mathbf{y}) - \bar{s}(\mathbf{x})\big)\Big) + O(\eta^2)p_{\theta;\mathbf{x}}(\mathbf{y}).$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{y}\sim p^{+}_{\theta;\mathbf{x}}}\big[r_m(\mathbf{x},\mathbf{y})\big] = \sum_{\mathbf{y}} p^{+}_{\theta;\mathbf{x}}(\mathbf{y}) r_m(\mathbf{x},\mathbf{y}) \\
&= \sum_{\mathbf{y}} p_{\theta;\mathbf{x}}(\mathbf{y}) r_m(\mathbf{x},\mathbf{y}) + \eta \sum_{\mathbf{y}} p_{\theta;\mathbf{x}}(\mathbf{y}) r_m(\mathbf{x},\mathbf{y})\big(s(\mathbf{x},\mathbf{y}) - \bar{s}(\mathbf{x})\big) + O(\eta^2)\sum_{\mathbf{y}} p_{\theta;\mathbf{x}}(\mathbf{y}) r_m(\mathbf{x},\mathbf{y}) \\
&= \mathbb{E}_{\mathbf{y}\sim p_{\theta;\mathbf{x}}}\big[r_m(\mathbf{x},\mathbf{y})\big] + \eta\Big(\mathbb{E}_{\mathbf{y}\sim p_{\theta;\mathbf{x}}}\big[r_m(\mathbf{x},\mathbf{y}) s(\mathbf{x},\mathbf{y})\big] - \mathbb{E}_{\mathbf{y}\sim p_{\theta;\mathbf{x}}}\big[r_m(\mathbf{x},\mathbf{y})\big]\mathbb{E}_{\mathbf{y}\sim p_{\theta;\mathbf{x}}}\big[s(\mathbf{x},\mathbf{y})\big]\Big) + O(\eta^2) \\
&= \mathbb{E}_{\mathbf{y}\sim p_{\theta;\mathbf{x}}}\big[r_m(\mathbf{x},\mathbf{y})\big] + \eta \mathrm{Cov}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big(r_m(\mathbf{x},\mathbf{y}), s(\mathbf{x},\mathbf{y})\big) + O(\eta^2).
\end{aligned}$$

The third equality follows since $r_m(\mathbf{x},\mathbf{y})$ is bounded, so the remainder term is $O(\eta^2)$. Finally, taking expectation over $\mathbf{x} \sim \mathcal{D}$ gives Equation (3). □

D.3 Proof of Lemma C.1

Proof. Define the deterministic one-dimensional function

$$g(t) \coloneqq r_m\big(\theta + td(\theta)\big),\ t \in [0,\eta].$$

By the fundamental theorem of calculus,

$$\begin{aligned}
r_m(\theta^+) - r_m(\theta) = g(\eta) - g(0) &= \int_0^\eta g'(t)dt \\
&= \int_0^\eta \big\langle d(\theta), \nabla r_m\big(\theta + td(\theta)\big)\big\rangle dt \\
&= \int_0^\eta \big\langle d(\theta), \nabla r_m(\theta)\big\rangle dt + \int_0^\eta \big\langle d(\theta), \nabla r_m\big(\theta + td(\theta)\big) - \nabla r_m(\theta)\big\rangle dt \\
&= \eta\big\langle \nabla r_m(\theta), d(\theta)\big\rangle + \int_0^\eta \big\langle \nabla r_m\big(\theta + td(\theta)\big) - \nabla r_m(\theta), d(\theta)\big\rangle dt.
\end{aligned}$$

For the second term, the Cauchy-Schwarz inequality implies

$$\big\langle \nabla r_m\big(\theta + td(\theta)\big) - \nabla r_m(\theta), d(\theta)\big\rangle \geq -\|\nabla r_m\big(\theta + td(\theta)\big) - \nabla r_m(\theta)\|\|d(\theta)\|.$$

By the L_m-Lipschitzness of ∇r_m,

$$\|\nabla r_m\big(\theta + td(\theta)\big) - \nabla r_m(\theta)\| \leq L_m\|\theta + td(\theta) - \theta\| = L_m t\|d(\theta)\|.$$

Therefore,

$$\int_0^\eta \big\langle \nabla r_m\big(\theta + td(\theta)\big) - \nabla r_m(\theta), d(\theta)\big\rangle dt \geq -\int_0^\eta L_m t\|d(\theta)\|^2 dt = -\frac{L_m}{2}\eta^2\|d(\theta)\|^2.$$

Combining the bounds yields

$$r_m(\theta^+) - r_m(\theta) \geq \eta\big\langle \nabla r_m(\theta), d(\theta)\big\rangle - \frac{L_m}{2}\eta^2\|d(\theta)\|^2,$$

which proves the lemma. □

D.4 Proof of Theorem C.2

Proof. Write the GRPO surrogate objective as

$$J(\theta) = \mathbb{E}\left[\sum_{k=1}^{K}\sum_{l=1}^{L_{\text{out}}} \min\left\{\rho_{k,l}(\theta)A_k(\mathbf{x}), \bar{\rho}_{k,l}(\theta)A_k(\mathbf{x})\right\}\right] - \beta\mathbb{E}\big[\mathrm{KL}(p_\theta\|p_{\text{ref}})\big] + \lambda\mathbb{E}\big[H(p_\theta)\big].$$

Fix $(\mathbf{x}, \mathbf{y}^{(1:K)})$ and a token index (k, l). Define

$$s_{k,l}(\theta) \coloneqq \min\left\{\rho_{k,l}(\theta)A_k(\mathbf{x}), \bar{\rho}_{k,l}(\theta)A_k(\mathbf{x})\right\}.$$

By the definition of tokenwise clipping, $s_{k,l}(\theta)$ equals $\rho_{k,l}(\theta)A_k(\mathbf{x})$ on the unclipped region and equals $\bar{\rho}_{k,l}(\theta)A_k(\mathbf{x})$ on the clipped region. On the clipped region, $\bar{\rho}_{k,l}(\theta)$ is constant in θ (equal to $1+\varepsilon$ or $1-\varepsilon$), so its gradient is zero. On the unclipped region,

$$\nabla s_{k,l}(\theta) = A_k(\mathbf{x})\nabla\rho_{k,l}(\theta) = A_k(\mathbf{x})\rho_{k,l}(\theta)\nabla\log p_\theta\big(y_l^{(k)} \mid \mathbf{x}, \mathbf{y}_{\le l-1}^{(k)}\big) = W_{k,l}(\theta)\phi_{k,l}(\theta),$$

where $W_{k,l}(\theta) = A_k(\mathbf{x})\rho_{k,l}(\theta)\mathbf{1}_{k,l}(\theta)$ encodes exactly the unclipped region. Therefore,

$$\nabla_\theta\mathbb{E}\Big[\sum_{k,l} s_{k,l}(\theta)\Big] = \mathbb{E}\Big[\sum_{k,l}\nabla s_{k,l}(\theta)\Big] = \mathbb{E}\Big[\sum_{k,l} W_{k,l}(\theta)\phi_{k,l}(\theta)\Big] = G(\theta).$$

For the regularizers,

$$\nabla_\theta\Big(-\beta\mathbb{E}[\mathrm{KL}(p_\theta\|p_{\text{ref}})] + \lambda\mathbb{E}[H(p_\theta)]\Big) = -\Big(\beta\nabla_\theta\mathbb{E}[\mathrm{KL}(p_\theta\|p_{\text{ref}})] - \lambda\nabla_\theta\mathbb{E}[H(p_\theta)]\Big) = -R(\theta).$$

Combining these two parts gives $\nabla J(\theta) = G(\theta) - R(\theta)$, hence

$$d_{\text{nat}}(\theta) = F(\theta)^{-1}\nabla J(\theta) = F(\theta)^{-1}\big(G(\theta) - R(\theta)\big).$$

By Lemma C.1, for each objective m and any direction $d(\theta)$,

$$r_m(\theta + \eta d(\theta)) - r_m(\theta) \ge \eta\nabla r_m(\theta)^\top d(\theta) - \frac{L_m}{2}\eta^2\|d(\theta)\|^2.$$

Setting $d(\theta) = d_{\text{nat}}(\theta)$ and using the definition of $d_{\text{nat}}(\theta)$ yields

$$\begin{aligned}\nabla r_m(\theta)^\top d_{\text{nat}}(\theta) &= \nabla r_m(\theta)^\top F(\theta)^{-1}\big(G(\theta) - R(\theta)\big)\\ &= \nabla r_m(\theta)^\top F(\theta)^{-1}G(\theta) - \nabla r_m(\theta)^\top F(\theta)^{-1}R(\theta).\end{aligned}$$

Thus the condition Equation (8) implies $\nabla r_m(\theta)^\top d_{\text{nat}}(\theta) \ge 0$ for all m. If $d_{\text{nat}}(\theta) = 0$, then $\theta^+ = \theta$ and $r_m(\theta^+) = r_m(\theta)$ trivially. Otherwise, if $\gamma_m(\theta) = \nabla r_m(\theta)^\top d_{\text{nat}}(\theta) > 0$ for all m, then

$$r_m(\theta^+) - r_m(\theta) \ge \eta\gamma_m(\theta) - \frac{L_m}{2}\eta^2\|d_{\text{nat}}(\theta)\|^2.$$

Choosing

$$0 < \eta \le \min_m \frac{2\gamma_m(\theta)}{L_m\|d_{\text{nat}}(\theta)\|^2}$$

ensures the right-hand side is nonnegative for every m, hence $r_m(\theta^+) \ge r_m(\theta)$ for all m. If $\min_m \gamma_m(\theta) > 0$, then taking η strictly smaller than the bound gives strict improvement for all m. □

D.5 Proof of Corollary C.3

Proof. Fix $\mathbf{x}$. For categorical softmax logits, for any $\mathbf{y}, \mathbf{z} \in \mathcal{X}^{L_{\text{out}}}$, the Jacobian satisfies

$$\frac{\partial p_\theta(\mathbf{y} \mid \mathbf{x})}{\partial\theta_{\mathbf{z}}} = p_\theta(\mathbf{y} \mid \mathbf{x})\big(\mathbf{1}\{\mathbf{y} = \mathbf{z}\} - p_\theta(\mathbf{z} \mid \mathbf{x})\big).$$

For each $\mathbf{z}$, differentiating the expectation yields

$$\begin{aligned}\frac{\partial}{\partial \theta_{\mathbf{z}}}\mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[w(\mathbf{x},\mathbf{y})\big] &= \sum_{\mathbf{y}} w(\mathbf{x},\mathbf{y})\frac{\partial p_\theta(\mathbf{y}\mid\mathbf{x})}{\partial \theta_{\mathbf{z}}}\\ &= \sum_{\mathbf{y}} w(\mathbf{x},\mathbf{y})p_\theta(\mathbf{y}\mid\mathbf{x})\big(\mathbf{1}\{\mathbf{y}=\mathbf{z}\}-p_\theta(\mathbf{z}\mid\mathbf{x})\big)\\ &= p_\theta(\mathbf{z}\mid\mathbf{x})w(\mathbf{x},\mathbf{z}) - p_\theta(\mathbf{z}\mid\mathbf{x})\sum_{\mathbf{y}} p_\theta(\mathbf{y}\mid\mathbf{x})w(\mathbf{x},\mathbf{y})\\ &= p_\theta(\mathbf{z}\mid\mathbf{x})\Big(w(\mathbf{x},\mathbf{z}) - \mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[w(\mathbf{x},\mathbf{y})\big]\Big).\end{aligned}$$

Therefore, we obtain its vector form

$$\nabla_\theta \mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[w(\mathbf{x},\mathbf{y})\big] = p_\theta(\cdot\mid\mathbf{x})\odot\Big(w(\mathbf{x},\cdot) - \mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[w(\mathbf{x},\mathbf{y})\big]\mathbf{1}\Big), \tag{9}$$

where $\odot$ denotes elementwise product. For the categorical softmax family, the Fisher matrix is

$$F(\theta) = \mathrm{diag}\big(p_\theta(\cdot\mid\mathbf{x})\big) - p_\theta(\cdot\mid\mathbf{x})p_\theta(\cdot\mid\mathbf{x})^\top.$$

For any vector v indexed by $\mathbf{y}$, the $\mathbf{y}$-th coordinate of $F(\theta)v$ is

$$\begin{aligned}\big[F(\theta)v\big]_{\mathbf{y}} &= \big[\mathrm{diag}(p_\theta(\cdot\mid\mathbf{x}))v\big]_{\mathbf{y}} - \big[p_\theta(\cdot\mid\mathbf{x})p_\theta(\cdot\mid\mathbf{x})^\top v\big]_{\mathbf{y}}\\ &= p_\theta(\mathbf{y}\mid\mathbf{x})v_{\mathbf{y}} - p_\theta(\mathbf{y}\mid\mathbf{x})\sum_{\mathbf{z}} p_\theta(\mathbf{z}\mid\mathbf{x})v_{\mathbf{z}}\\ &= p_\theta(\mathbf{y}\mid\mathbf{x})\Big(v_{\mathbf{y}} - \mathbb{E}_{\mathbf{z}\sim p_\theta(\cdot|\mathbf{x})}[v_{\mathbf{z}}]\Big).\end{aligned}$$

Now take

$$v \;=\; w(\mathbf{x},\cdot) - \mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[w(\mathbf{x},\mathbf{y})\big]\mathbf{1}.$$

Then

$$\begin{aligned}\mathbb{E}_{\mathbf{z}\sim p_\theta(\cdot|\mathbf{x})}[v_{\mathbf{z}}] &= \mathbb{E}_{\mathbf{z}\sim p_\theta(\cdot|\mathbf{x})}\big[w(\mathbf{x},\mathbf{z})\big] - \mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[w(\mathbf{x},\mathbf{y})\big]\cdot\mathbb{E}_{\mathbf{z}\sim p_\theta(\cdot|\mathbf{x})}[1]\\ &= 0,\end{aligned}$$

so the general coordinate formula reduces to

$$\big[F(\theta)v\big]_{\mathbf{y}} = p_\theta(\mathbf{y}\mid\mathbf{x})v_{\mathbf{y}}.$$

Equivalently,

$$F(\theta)v = p_\theta(\cdot\mid\mathbf{x})\odot v = p_\theta(\cdot\mid\mathbf{x})\odot\Big(w(\mathbf{x},\cdot) - \mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[w(\mathbf{x},\mathbf{y})\big]\mathbf{1}\Big).$$

Comparing with the expression for $\nabla_\theta\mathbb{E}[w]$ from Equation (9) shows that this particular v satisfies

$$F(\theta)v = \nabla_\theta\mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[w(\mathbf{x},\mathbf{y})\big].$$

By definition, any vector $d_{\mathrm{nat}}(\theta)$ satisfying

$$F(\theta)d_{\mathrm{nat}}(\theta) = \nabla_\theta\mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[w(\mathbf{x},\mathbf{y})\big]$$

is a valid natural gradient direction (the solution is not unique because $F(\theta)\mathbf{1} = 0$). Therefore, one convenient and valid choice in the space $\mathbf{1}^\top d_{\mathrm{nat}}(\theta) = 0$ is

$$d_{\mathrm{nat}}(\theta) = w(\mathbf{x},\cdot) - \mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[w(\mathbf{x},\mathbf{y})\big]\mathbf{1}. \tag{10}$$

Take $\theta^+ = \theta + \eta d_{\mathrm{nat}}(\theta)$ with learning rate $\eta > 0$. A first-order Taylor expansion gives

$$p_{\theta^+}(\mathbf{y}\mid\mathbf{x}) - p_\theta(\mathbf{y}\mid\mathbf{x}) = \eta\sum_{\mathbf{z}}\frac{\partial p_\theta(\mathbf{y}\mid\mathbf{x})}{\partial\theta_{\mathbf{z}}}d_{\mathrm{nat}}(\theta)_{\mathbf{z}} + O(\eta^2).$$

Substituting the Jacobian formula and simplifying,

$$\begin{aligned}p_{\theta^+}(\mathbf{y}\mid\mathbf{x}) - p_\theta(\mathbf{y}\mid\mathbf{x}) &= \eta p_\theta(\mathbf{y}\mid\mathbf{x})\Big(d_{\mathrm{nat}}(\theta)_{\mathbf{y}} - \sum_{\mathbf{z}} p_\theta(\mathbf{z}\mid\mathbf{x})d_{\mathrm{nat}}(\theta)_{\mathbf{z}}\Big) + O(\eta^2)\\ &= \eta p_\theta(\mathbf{y}\mid\mathbf{x})\Big(d_{\mathrm{nat}}(\theta)_{\mathbf{y}} - \mathbb{E}_{\mathbf{z}\sim p_\theta(\cdot|\mathbf{x})}\big[d_{\mathrm{nat}}(\theta)_{\mathbf{z}}\big]\Big) + O(\eta^2).\end{aligned}$$

By Equation (10),

$$\mathbb{E}_{\mathbf{z}\sim p_\theta(\cdot|\mathbf{x})}\big[d_{\mathrm{nat}}(\theta)_{\mathbf{z}}\big] = \mathbb{E}_{\mathbf{z}\sim p_\theta(\cdot|\mathbf{x})}\big[w(\mathbf{x},\mathbf{z})\big] - \mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[w(\mathbf{x},\mathbf{y})\big] = 0,$$

and hence

$$p_{\theta^+}(\mathbf{y}\mid\mathbf{x}) - p_\theta(\mathbf{y}\mid\mathbf{x}) = \eta p_\theta(\mathbf{y}\mid\mathbf{x})\Big(w(\mathbf{x},\mathbf{y}) - \mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[w(\mathbf{x},\mathbf{y})\big]\Big) + O(\eta^2).$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{y}\sim p_{\theta^+}(\cdot|\mathbf{x})}\big[r_m(\mathbf{x},\mathbf{y})\big] - \mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[r_m(\mathbf{x},\mathbf{y})\big] \\
&\quad= \sum_{\mathbf{y}}\Big(p_{\theta^+}(\mathbf{y}\mid\mathbf{x}) - p_\theta(\mathbf{y}\mid\mathbf{x})\Big) r_m(\mathbf{x},\mathbf{y}) \\
&\quad= \eta\sum_{\mathbf{y}} p_\theta(\mathbf{y}\mid\mathbf{x})\Big(w(\mathbf{x},\mathbf{y}) - \mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[w(\mathbf{x},\mathbf{y})\big]\Big) r_m(\mathbf{x},\mathbf{y}) + O(\eta^2) \\
&\quad= \eta\Big(\mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[r_m(\mathbf{x},\mathbf{y})w(\mathbf{x},\mathbf{y})\big] - \mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[r_m(\mathbf{x},\mathbf{y})\big]\mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big[w(\mathbf{x},\mathbf{y})\big]\Big) + O(\eta^2) \\
&\quad= \eta\mathrm{Cov}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}\big(r_m(\mathbf{x},\mathbf{y}), w(\mathbf{x},\mathbf{y})\big) + O(\eta^2),
\end{aligned}$$

which proves the claim. □

D.6 Proof of Corollary C.4

Proof. Fix an objective m and write $F \coloneqq F(\theta)$, $R \coloneqq R(\theta)$, $G^{\mathrm{unclip}} \coloneqq G^{\mathrm{unclip}}(\theta)$, and $G^{\mathrm{clip}} \coloneqq G^{\mathrm{clip}}(\theta)$. By definition,

$$\gamma_m^{\mathrm{unclip}}(\theta) = \nabla r_m(\theta)^\top F^{-1}\big(G^{\mathrm{unclip}} - R\big).$$

Hence we can decompose the clipped first-order margin as

$$\begin{aligned}
\nabla r_m(\theta)^\top F^{-1}\big(G^{\mathrm{clip}} - R\big) &= \nabla r_m(\theta)^\top F^{-1}\big(G^{\mathrm{unclip}} - R\big) - \nabla r_m(\theta)^\top F^{-1}\big(G^{\mathrm{unclip}} - G^{\mathrm{clip}}\big) \\
&= \gamma_m^{\mathrm{unclip}}(\theta) - \nabla r_m(\theta)^\top F^{-1}\big(G^{\mathrm{unclip}} - G^{\mathrm{clip}}\big).
\end{aligned}$$

Now insert $F^{-1/2}$ and apply Cauchy-Schwarz inequality:

$$\begin{aligned}
\Big|\nabla r_m(\theta)^\top F^{-1}\big(G^{\mathrm{unclip}} - G^{\mathrm{clip}}\big)\Big| &= \Big|\big(F^{-1/2}\nabla r_m(\theta)\big)^\top\big(F^{-1/2}(G^{\mathrm{unclip}} - G^{\mathrm{clip}})\big)\Big| \\
&\le \big\|F^{-1/2}\nabla r_m(\theta)\big\| \cdot \big\|F^{-1/2}\big(G^{\mathrm{unclip}} - G^{\mathrm{clip}}\big)\big\|.
\end{aligned}$$

Combining with the decomposition yields the margin bound

$$\nabla r_m(\theta)^\top F^{-1}\big(G^{\mathrm{clip}} - R\big) \ge \gamma_m^{\mathrm{unclip}}(\theta) - \big\|F^{-1/2}\nabla r_m(\theta)\big\| \cdot \big\|F^{-1/2}\big(G^{\mathrm{unclip}} - G^{\mathrm{clip}}\big)\big\|.$$

Therefore, if $\gamma_m^{\mathrm{unclip}}(\theta) \ge \kappa_m > 0$ and

$$\big\|F^{-1/2}\big(G^{\mathrm{unclip}} - G^{\mathrm{clip}}\big)\big\| \le \frac{\kappa_m}{\big\|F^{-1/2}\nabla r_m(\theta)\big\|},$$

then $\nabla r_m(\theta)^\top F^{-1}(G^{\mathrm{clip}} - R) \ge 0$, which proves the claim. □

D.7 Proof of Theorem D.6

We first restate, in full, the assumptions and the theorem.

Assumption D.3 (Bounded score and unique optimal completion)**.** There exists $B > 0$ such that $|s(\mathbf{x},\mathbf{y})| \le B$ for all $\mathbf{y} \in \mathcal{X}^{L_{\mathrm{out}}}$. Moreover, there exists a unique maximizer $\mathbf{y}^* = \arg\max_{\mathbf{y}} s(\mathbf{x},\mathbf{y})$ and a margin $\Delta_s > 0$ such that $s(\mathbf{x},\mathbf{y}^*) - s(\mathbf{x},\mathbf{y}) \ge \Delta_s$ for all $\mathbf{y} \ne \mathbf{y}^*$.

Assumption D.4 (Non-saturation for suboptimal policy)**.** There exists $\epsilon \in (0,1)$ such that for every suboptimal parameter θ (i.e., $V(\mathbf{x};\theta) < V(\mathbf{x};\theta^*)$), token probabilities are bounded away from 1: $p_\theta\big(y_l \mid \mathbf{x}, \mathbf{y}_{\le l-1}\big) \le 1-\epsilon,\ \forall l \in \{1,\dots,L_{\mathrm{out}}\},\ \forall y_l \in \mathcal{X}$.

Assumption D.5 (Aligned token gradients)**.** There exists $c \in (0, 1]$ such that for any $\mathbf{y}$ and any positions $l, k \in \{1, \dots, L_{\text{out}}\}$ with $v_l(\mathbf{x}, \mathbf{y}; \theta) \neq 0$ and $v_k(\mathbf{x}, \mathbf{y}; \theta) \neq 0$,

$$\cos\big(v_l(\mathbf{x}, \mathbf{y}; \theta), v_k(\mathbf{x}, \mathbf{y}; \theta)\big) \geq c, \text{ where } v_l(\mathbf{x}, \mathbf{y}; \theta) \coloneqq J^\top_{f(\mathbf{x}, \mathbf{y}_{\leq l-1}; \theta)}\Big(\mathbf{e}_{y_l} - p_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{\leq l-1})\Big) \in \mathbb{R}^n.$$

$v_l(\mathbf{x}, \mathbf{y}; \theta)$ is the token-level logit-gradient contribution with f the logit map and J_f its Jacobian. These assumptions have clear interpretations in our setting. Theorem D.3 is natural because every reward in the paper lies in $[0, 1]$. Theorem D.4 matches practical RFT setups where KL regularization prevents token probabilities from saturating at 1. Theorem D.5 is stronger and model-dependent but has a direct geometric meaning: token-level gradient contributions along a favorable trajectory should be sufficiently aligned, so that they reinforce rather than cancel each other.

Theorem D.6. *(μ-PL condition of multi-objective alignment) For parameters $\theta \in \mathbb{R}^n$ and input $\mathbf{x} \in \mathcal{X}^{L_{\text{in}}}$, define the scalarized multi-objective value function $V(\mathbf{x}, \theta)$ as in Equation* (1)*. If the scalarization function s and policy θ meet the Assumptions D.3-D.5, then it holds that:*

$$\frac{1}{2}\|\nabla_\theta V(\mathbf{x}; \theta)\|^2 \geq \mu(V(\mathbf{x}; \theta^*) - V(\mathbf{x}; \theta)), \text{ with } \mu = \frac{1}{2B}\Big(\frac{p_\theta(\mathbf{y}^* \mid \mathbf{x})}{1 - p_\theta(\mathbf{y}^* \mid \mathbf{x})} s(\mathbf{x}, \mathbf{y}^*)\gamma - 2B\sigma_{\max}\Big).$$

$\gamma = \sqrt{cL_{\text{out}}\sigma^2_{\min}\epsilon^2 \frac{|\mathcal{X}^{L_{\text{out}}}|}{|\mathcal{X}^{L_{\text{out}}}|-1}}$. $\sigma_{\max}$ *and* $\sigma_{\min}$ *are the largest and smallest singular values of* $J_{f(\mathbf{x}, \mathbf{y}_{\leq l-1}; \theta)}$.

We first recall the PL condition for maximization.

Definition D.7. Let $V : \mathbb{R}^n \to \mathbb{R}$ be a differentiable objective function that we aim to *maximize*. Define the set of global maximizers $\Theta^* \coloneqq \arg\max_{\theta \in \mathbb{R}^n} V(\theta)$, and assume $\Theta^* \neq \emptyset$. We say that V satisfies the μ-PL condition with parameter $\mu > 0$ on a set $\Omega \subseteq \mathbb{R}^n$ if

$$\frac{1}{2}\big\|\nabla_\theta V(\theta)\big\|^2 \geq \mu\big(V(\theta^*) - V(\theta)\big), \ \forall \theta \in \Omega, \ \theta^* \in \Theta^* \tag{11}$$

In words, the PL condition ties gradient magnitude to global suboptimality, such that any point with a small gradient norm must have an objective value close to the global maximum, even though V need not be convex. We now prove that the scalarized RFT objective satisfies this condition under Assumptions D.3–D.5.

Proof. We first differentiate $V(\mathbf{x}; \theta)$ with respect to θ, using the log-derivative trick gives:

$$\begin{aligned}
\nabla_\theta V(\mathbf{x}; \theta) &= \mathbb{E}_{\mathbf{y} \sim p_\theta(\cdot \mid \mathbf{x})}\Big[s(\mathbf{x}, \mathbf{y}) \nabla_\theta \ln p_\theta(\mathbf{y} \mid \mathbf{x})\Big] \\
&= \mathbb{E}_{\mathbf{y} \sim p_\theta(\cdot \mid \mathbf{x})}\Big[s(\mathbf{x}, \mathbf{y}) \sum_{l=1}^{L_{\text{out}}} \nabla_\theta \ln p_\theta(y_l \mid \mathbf{x}, \mathbf{y}_{\leq l-1})\Big] \\
&= \mathbb{E}_{\mathbf{y} \sim p_\theta(\cdot \mid \mathbf{x})}\Big[s(\mathbf{x}, \mathbf{y}) \sum_{l=1}^{L_{\text{out}}} \nabla_\theta \ln \operatorname{softmax}(f(\mathbf{x}, \mathbf{y}_{\leq l-1}; \theta))_{y_l}\Big] \\
&= \sum_{\mathbf{y} \in \mathcal{X}^{L_{\text{out}}}} p_\theta(\mathbf{y} \mid \mathbf{x}) s(\mathbf{x}, \mathbf{y}) \sum_{l=1}^{L_{\text{out}}} J^\top_{f(\mathbf{x}, \mathbf{y}_{\leq l-1}; \theta)}(\mathbf{e}_{y_l} - p_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{\leq l-1})).
\end{aligned}$$

where $J_{f(\mathbf{x}, \mathbf{y}_{\leq l-1}; \theta)}$ is the Jacobian of $f(\mathbf{x}, \mathbf{y}_{\leq l-1}; \theta)$ with respect to θ. $\mathbf{e}_{y_l}$ denotes the y_l'th standard basis vector whose y_l'th entry is 1 and all other entries are 0. $p_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{\leq l-1}) = \operatorname{softmax}(f(\mathbf{x}, \mathbf{y}_{\leq l-1}; \theta)) \in \mathbb{R}^{|\mathcal{X}|}$ returns the next-token (at position l) probability distribution. Let $\phi(\mathbf{x}, \mathbf{y}; \theta) \coloneqq \sum_{l=1}^{L_{\text{out}}} J^\top_{f(\mathbf{x}, \mathbf{y}_{\leq l-1}; \theta)}(\mathbf{e}_{y_l} - p_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{\leq l-1}))$ and $\sigma_{\max}$ be the largest singular value of $J_{f(\mathbf{x}, \mathbf{y}_{\leq l-1}; \theta)}$. By definition, we know $\sigma_{\max} = \max_{l \in \{1, 2, \cdots, L_{\text{out}}\}} \|J_{f(\mathbf{x}, \mathbf{y}_{\leq l-1}; \theta)}\|$. Obtaining the upper bound of $\phi(\mathbf{x}, \mathbf{y}; \theta)$ is nontrivial:

$$\|\phi(\mathbf{x}, \mathbf{y}; \theta)\| \leq \|J^\top_{f(\mathbf{x}, \mathbf{y}_{\leq l-1}; \theta)}\|_2 \cdot \|(\mathbf{e}_{y_l} - p_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{\leq l-1}))\| \leq 2\sigma_{\max}$$

because $\|(\mathbf{e}_{y_l} - p_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{\leq l-1}))\| \leq \|(\mathbf{e}_{y_l} - p_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{\leq l-1}))\|_1 \leq 2$ and $\|J^\top_{f(\mathbf{x}, \mathbf{y}_{\leq l-1}; \theta)}\|_2 = \sigma_{\max}$ by definition. However, we cannot lower bound $\phi(\mathbf{x}, \mathbf{y}; \theta)$ without extra structure. So we first assume

token probabilities are bounded away from 1 for all suboptimal θ in Assumption D.4, consistent with practical RFT setups where KL or entropy regularization prevents probabilities from overfitting to 1 during training. Assumption D.5 requires all nonzero per-token contributions v_l are positively aligned in parameter space and their cosine similarity is uniformly bounded away from zero by c. With these assumptions about the structure of the policy function, we can derive the following lemmas that will be used in the proof later.

Lemma D.8. *Under Assumptions D.4 and D.5, we have*

$$\Big\| \sum_{l=1}^{L_{\text{out}}} J^{\top}_{f(\mathbf{x},\mathbf{y}_{\le l-1};\theta)} (\mathbf{e}_{y_l} - p_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{\le l-1})) \Big\|^2 \ge c L_{\text{out}} \sigma_{\min}^2 \epsilon^2 \frac{|\mathcal{X}^{L_{\text{out}}}|}{|\mathcal{X}^{L_{\text{out}}}| - 1} \tag{12}$$

Remark D.9. The left-hand side of Equation (12) quantifies the squared norm of the summed token-level policy gradient contributions along an output sequence of length L_{out}, and thus measures the overall strength of the gradient signal induced by that sequence. The lower bound shows that this signal scales with the alignment constant c, the smallest singular value $\sigma_{\min}$ of the logit Jacobian, the sequence length L_{out}, and the probability gap ϵ. Here, ϵ lower bounds $1 - p_\theta(y_l \mid \mathbf{x}, \mathbf{y}_{\le l-1})$ and captures the policy's uncertainty in predicting the next token. Smaller ϵ corresponds to higher confidence and more likely token sampling, but also leads to weaker gradient signals.

Proof of Lemma D.8. We lower bound the L2-norm of the vector $\mathbf{e}_{y_l} - p_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{\le l-1})$:

$$\begin{aligned}
\|\mathbf{e}_{y_l} - p_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{\le l-1})\|^2 &= \big(1 - p_\theta(y_l \mid \mathbf{x}, \mathbf{y}_{\le l-1})\big)^2 + \sum_{y' \ne y_l, y' \in \mathcal{X}} p_\theta^2(y' \mid \mathbf{x}, \mathbf{y}_{\le l-1}) \\
&\ge \big(1 - p_\theta(y_l \mid \mathbf{x}, \mathbf{y}_{\le l-1})\big)^2 + \frac{\big(\sum_{y' \ne y_l, y' \in \mathcal{X}} p_\theta(y' \mid \mathbf{x}, \mathbf{y}_{\le l-1})\big)^2}{|\mathcal{X}^{L_{\text{out}}}| - 1} \\
&= \big(1 - p_\theta(y_l \mid \mathbf{x}, \mathbf{y}_{\le l-1})\big)^2 + \frac{\big(1 - p_\theta(y_l \mid \mathbf{x}, \mathbf{y}_{\le l-1})\big)^2}{|\mathcal{X}^{L_{\text{out}}}| - 1} \\
&= \frac{|\mathcal{X}^{L_{\text{out}}}|}{|\mathcal{X}^{L_{\text{out}}}| - 1} \big(1 - p_\theta(y_l \mid \mathbf{x}, \mathbf{y}_{\le l-1})\big)^2 \\
&\ge \epsilon^2 \frac{|\mathcal{X}^{L_{\text{out}}}|}{|\mathcal{X}^{L_{\text{out}}}| - 1}.
\end{aligned} \tag{13}$$

The first inequality is derived from the Cauchy-Schwarz inequality in Euclidean space $(\sum p_\theta)^2 \le (\sum p_\theta^2)(\sum 1^2)$, and the last inequality is derived from the Assumption D.4. Then we have:

$$\begin{aligned}
\Big\| \sum_{l=1}^{L_{\text{out}}} J^{\top}_{f(\mathbf{x},\mathbf{y}_{\le l-1};\theta)} (\mathbf{e}_{y_l} - p_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{\le l-1})) \Big\|^2 &= \sum_{l=1}^{L_{\text{out}}} \Big\| v_l \Big\|^2 + 2 \sum_{l<k} \langle v_l, v_k \rangle \\
&\ge \sum_{l=1}^{L_{\text{out}}} \Big\| v_l \Big\|^2 + 2c \sum_{l<k} \|v_l\| \|v_k\| \\
&\ge c \Big(\sum_{l=1}^{L_{\text{out}}} \|v_l\| \Big)^2 \\
&\ge c \sigma_{\min}^2 \Big(\sum_{l=1}^{L_{\text{out}}} \|\mathbf{e}_{y_l} - p_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{\le l-1})\| \Big)^2 \\
&\ge c L_{\text{out}} \sigma_{\min}^2 \epsilon^2 \frac{|\mathcal{X}^{L_{\text{out}}}|}{|\mathcal{X}^{L_{\text{out}}}| - 1}.
\end{aligned}$$

Here, $\sigma_{\min}$ denotes the smallest singular value of the Jacobian $J_{f(\mathbf{x},\mathbf{y}_{\le l-1};\theta)}$. The second inequality follows from the Assumption D.5 and factors out c. The third inequality applies the singular value bound $\|J^{\top}_{f(\mathbf{x},\mathbf{y}_{\le l-1};\theta)} u\| \ge \sigma_{\min} \|u\|, \ \forall u$. The final inequality uses the result from Equation (13). □

Lemma D.10. *Let Assumption D.3 hold and $\mathbf{y}^*$ is the unique maximizer of the scalarized reward, $\mathbf{y}^* \in \arg\max_{\mathbf{y}} s(\mathbf{x}, \mathbf{y})$. Assume the policy class is rich enough to realize a deterministic policy on $\mathbf{x}$ that always outputs $\mathbf{y}^*$, i.e., there exists a parameter vector θ^* such that*

$$p_{\theta^*}(\mathbf{y} \mid \mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{y} = \mathbf{y}^*, \\ 0, & \text{otherwise}. \end{cases}$$

Then θ^* is an optimal parameter for the scalarized value at $\mathbf{x}$, and the optimal value becomes $V(\mathbf{x}; \theta^*) = \sum_{\mathbf{y} \in \mathcal{X}^{L_{\text{out}}}} p_{\theta^*}(\mathbf{y} \mid \mathbf{x}) s(\mathbf{x}, \mathbf{y}) = s(\mathbf{x}, \mathbf{y}^*)$. For any θ, we have:

$$V(\mathbf{x}; \theta^*) - V(\mathbf{x}; \theta) \le 2B\big(1 - p_\theta(\mathbf{y}^* \mid \mathbf{x})\big).$$

Remark D.11. Lemma D.10 shows that the value gap is Lipschitz in the probability gap on $\mathbf{y}^*$. Every unit of probability that fails to go to the optimal sequence can hurt the value by at most $2B$. Equivalently, driving the model towards near-deterministic predictions on $\mathbf{y}^*$, $p_\theta(\mathbf{y}^* \mid \mathbf{x}) \approx 1$, is both necessary and sufficient (up to the factor $2B$) to make the scalarized value nearly optimal. This connects convergence in value directly to how sharply the policy concentrates on the optimal sequence.

Proof of Lemma D.10. We can write $V(\mathbf{x}; \theta^*) - V(\mathbf{x}; \theta)$ as follows:

$$\begin{aligned}
V(\mathbf{x}; \theta^*) - V(\mathbf{x}; \theta) &= s(\mathbf{x}, \mathbf{y}^*) - \sum_{\mathbf{y} \in \mathcal{X}^{L_{\text{out}}}} p_\theta(\mathbf{y} \mid \mathbf{x}) s(\mathbf{x}, \mathbf{y}) \\
&= s(\mathbf{x}, \mathbf{y}^*) - \big(p_\theta(\mathbf{y}^* \mid \mathbf{x}) s(\mathbf{x}, \mathbf{y}^*) + \sum_{\mathbf{y} \neq \mathbf{y}^*} p_\theta(\mathbf{y} \mid \mathbf{x}) s(\mathbf{x}, \mathbf{y})\big) \\
&= \big(1 - p_\theta(\mathbf{y}^* \mid \mathbf{x})\big) s(\mathbf{x}, \mathbf{y}^*) - \sum_{\mathbf{y} \neq \mathbf{y}^*} p_\theta(\mathbf{y} \mid \mathbf{x}) s(\mathbf{x}, \mathbf{y}) \\
&\le \big(1 - p_\theta(\mathbf{y}^* \mid \mathbf{x})\big) s(\mathbf{x}, \mathbf{y}^*) - \min_{\mathbf{y} \neq \mathbf{y}^*} s(\mathbf{x}, \mathbf{y}) \sum_{\mathbf{y} \neq \mathbf{y}^*} p_\theta(\mathbf{y} \mid \mathbf{x}) \\
&= \big(1 - p_\theta(\mathbf{y}^* \mid \mathbf{x})\big) s(\mathbf{x}, \mathbf{y}^*) - \min_{\mathbf{y} \neq \mathbf{y}^*} s(\mathbf{x}, \mathbf{y}) \big(1 - p_\theta(\mathbf{y}^* \mid \mathbf{x})\big) \\
&= \big(s(\mathbf{x}, \mathbf{y}^*) - \min_{\mathbf{y} \neq \mathbf{y}^*} s(\mathbf{x}, \mathbf{y})\big) \big(1 - p_\theta(\mathbf{y}^* \mid \mathbf{x})\big) \\
&\le 2B\big(1 - p_\theta(\mathbf{y}^* \mid \mathbf{x})\big).
\end{aligned}$$

The last inequality follows from the Assumption D.3 that $|s(\mathbf{x}, \mathbf{y})| \le B$ and $\|s(\mathbf{x}, \mathbf{y}^*) - \min_{\mathbf{y} \neq \mathbf{y}^*} s(\mathbf{x}, \mathbf{y})\| \le \|s(\mathbf{x}, \mathbf{y}^*) - \min_{\mathbf{y} \neq \mathbf{y}^*} s(\mathbf{x}, \mathbf{y})\|_1 \le 2B$. □

Defining a direction vector:

$$u(\mathbf{x}, \mathbf{y}^*; \theta) \coloneqq \frac{\phi(\mathbf{x}, \mathbf{y}^*; \theta)}{\|\phi(\mathbf{x}, \mathbf{y}^*; \theta)\|} \coloneqq \frac{\sum_{l=1}^{L_{\text{out}}} J^\top_{f(\mathbf{x}, \mathbf{y}_{\le l-1}; \theta)} \big(\mathbf{e}_{y_l} - p_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{\le l-1})\big)}{\big\| \sum_{l=1}^{L_{\text{out}}} J^\top_{f(\mathbf{x}, \mathbf{y}_{\le l-1}; \theta)} \big(\mathbf{e}_{y_l} - p_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{\le l-1})\big) \big\|},$$

which is the normalized log-probability gradient of the optimal trajectory. We first discuss its inner product with $\phi(\mathbf{x}, \mathbf{y}; \theta)$ which will be used for the proof later.

$$\langle \phi(\mathbf{x}, \mathbf{y}; \theta), u(\mathbf{x}, \mathbf{y}^*; \theta) \rangle \begin{cases} = \|\phi(\mathbf{x}, \mathbf{y}^*; \theta)\| \ge \sqrt{c L_{\text{out}} \sigma_{\min}^2 \epsilon^2 \frac{|\mathcal{X}^{L_{\text{out}}}|}{|\mathcal{X}^{L_{\text{out}}}| - 1}}, & \text{if } \mathbf{y} = \mathbf{y}^*, \\ \le \|\phi(\mathbf{x}, \mathbf{y}; \theta)\| \le 2\sigma_{\max}, & \forall \mathbf{y}. \end{cases} \tag{14}$$

Then the directional derivative of $V(\mathbf{x}; \theta)$ along $u(\mathbf{x}, \mathbf{y}^*; \theta)$ is:

$$\langle \nabla_\theta V(\mathbf{x}; \theta), u(\mathbf{x}, \mathbf{y}^*; \theta) \rangle = \mathbb{E}_{\mathbf{y} \sim p_\theta(\cdot \mid \mathbf{x})} \Big[s(\mathbf{x}, \mathbf{y}) \langle \phi(\mathbf{x}, \mathbf{y}; \theta), u(\mathbf{x}, \mathbf{y}^*; \theta) \rangle \Big].$$

Extracting $\mathbf{y}^*$ out yields:

$$\begin{aligned}
\langle \nabla_\theta V(\mathbf{x}; \theta), u(\mathbf{x}, \mathbf{y}^*; \theta) \rangle = &\underbrace{p_\theta(\mathbf{y}^* \mid \mathbf{x}) s(\mathbf{x}, \mathbf{y}^*) \langle \phi(\mathbf{x}, \mathbf{y}^*; \theta), u(\mathbf{x}, \mathbf{y}^*; \theta) \rangle}_{(I)} + \\
&\underbrace{\sum_{\mathbf{y} \neq \mathbf{y}^*} p_\theta(\mathbf{y} \mid \mathbf{x}) s(\mathbf{x}, \mathbf{y}) \langle \phi(\mathbf{x}, \mathbf{y}; \theta), u(\mathbf{x}, \mathbf{y}^*; \theta) \rangle}_{(II)}.
\end{aligned} \tag{15}$$

We lower bound (I) and (II) separately. Starting with (I), by Equation (14), we know that:

$$(I) \geq p_\theta(\mathbf{y}^* \mid \mathbf{x}) s(\mathbf{x}, \mathbf{y}^*) \sqrt{c L_{\text{out}} \sigma_{\min}^2 \epsilon^2 \frac{|\mathcal{X}^{L_{\text{out}}}|}{|\mathcal{X}^{L_{\text{out}}}| - 1}}.$$

Recalling that $|s(\mathbf{x}, \mathbf{y})| \leq B$ from Assumption D.3 and the property $\langle \phi(\mathbf{x}, \mathbf{y}; \theta), u(\mathbf{x}, \mathbf{y}^*; \theta) \rangle \leq 2\sigma_{\max}$, we get:

$$(II) \geq -2B\sigma_{\max} \sum_{\mathbf{y} \neq \mathbf{y}^*} p_\theta(\mathbf{y} \mid \mathbf{x}) = -2B\sigma_{\max}(1 - p_\theta(\mathbf{y}^* \mid \mathbf{x})).$$

For (II), the worst case happens when, for every $\mathbf{y} \neq \mathbf{y}^*$, the term's contribution to the directional derivative along u is as adverse as allowed by the assumed bounds. Plugging the above two inequalities into Equation (15), rearranging the equality, it follows that:

$$\langle \nabla_\theta V(\mathbf{x}; \theta), u(\mathbf{x}, \mathbf{y}^*; \theta) \rangle \geq s(\mathbf{x}, \mathbf{y}^*)\gamma - \big(s(\mathbf{x}, \mathbf{y}^*)\gamma + 2B\sigma_{\max}\big)\big(1 - p_\theta(\mathbf{y}^* \mid \mathbf{x})\big),$$
$$\text{with } \gamma = \sqrt{c L_{\text{out}} \sigma_{\min}^2 \epsilon^2 \frac{|\mathcal{X}^{L_{\text{out}}}|}{|\mathcal{X}^{L_{\text{out}}}| - 1}}. \tag{16}$$

The constant μ quantifies how "sharp" the scalarized landscape $V(\mathbf{x}; \theta)$ is around its maximizer θ^*: it is large when the policy places large mass on the optimal completion $\mathbf{y}^*$, when the optimal score $s(\mathbf{x}, \mathbf{y}^*)$ is large, and when the logit map is well-conditioned (large γ, which summarizes the Jacobian and non-saturation constants). Conversely, μ becomes non-positive when the aligned ascent signal from the optimal completion is weak (small $p_\theta(\mathbf{y}^* \mid \mathbf{x})$ or small γ) or when the logit map is highly skewed (large $\sigma_{\max}/\sigma_{\min}$), so the identical scalarized update can converge on one model and stall on another.

Intuitively, this inequality shows that when the policy is close to the optimal θ^* and $\mathbf{y}^*$ is the (unique) optimal trajectory maximizing the scalarized reward, the directional derivative along u admits a large lower bound and thus yields a strong, well-aligned gradient update that pushes additional probability mass toward $\mathbf{y}^*$.

Recall that we have upper bounded the value function gap by the sequence probability gap in Lemma D.10, where we assume $p_{\theta^*}(\mathbf{y}^* \mid \mathbf{x}) = 1$. We are close to the result and want a final inequality that looks like:

$$\|\nabla_\theta V(\mathbf{x}; \theta)\| \geq \langle \nabla_\theta V(\mathbf{x}; \theta), u(\mathbf{x}, \mathbf{y}^*; \theta) \rangle \geq \mu'\big(p_{\theta^*}(\mathbf{y}^* \mid \mathbf{x}) - p_\theta(\mathbf{y}^* \mid \mathbf{x})\big) \geq \mu\big(V(\mathbf{x}; \theta^*) - V(\mathbf{x}; \theta)\big). \tag{17}$$

The first inequality follows from the Cauchy-Schwarz inequality applied to the unit vector and the last inequality comes from Lemma D.10. So the only task left here is to get a μ' that satisfies the second inequality for all θ. Combining with Equation (16), we have:

$$\mu' \leq \frac{p_\theta(\mathbf{y}^* \mid \mathbf{x})}{1 - p_\theta(\mathbf{y}^* \mid \mathbf{x})} s(\mathbf{x}, \mathbf{y}^*)\gamma - 2B\sigma_{\max}.$$

Because the function $\frac{p}{1-p}$ is increasing on $[0, 1)$, a sufficient and tight choice for μ' is:

$$\mu' \coloneqq \frac{p_\theta(\mathbf{y}^* \mid \mathbf{x})}{1 - p_\theta(\mathbf{y}^* \mid \mathbf{x})} s(\mathbf{x}, \mathbf{y}^*)\gamma - 2B\sigma_{\max}$$

Therefore, the final μ-PL condition follows from Equation (17), and combining it with the bound from Lemma D.10 leads to:

$$\mu = \frac{1}{2B}\Big(\frac{p_\theta(\mathbf{y}^* \mid \mathbf{x})}{1 - p_\theta(\mathbf{y}^* \mid \mathbf{x})} s(\mathbf{x}, \mathbf{y}^*)\gamma - 2B\sigma_{\max}\Big) \tag{18}$$

□

E ALGORITHMS

We provide pseudocode for algorithms that, to our knowledge, have not been previously applied to LLM alignment. Throughout the paper we report the smoothed variant of Tchebycheff scalarization rather than the classical one (Bowman, 1976), because it is more stable and stronger in our experiments. For dynamic weighting (Lu et al., 2026), PAMA (He & Maghsudi, 2026), and common linear scalarization, we follow their original implementations.

Algorithm 1 COVER on top of GRPO

Inputs: Objectives $\{r_m\}_{m=1}^{M}$, scalarization weights $\{\lambda_m\}$, COVER targets $\{c_m^*\}$, EMA rate $\tau \in (0, 1]$, weight learning rate η_λ.
Initialize $u_m \leftarrow \log \lambda_m$.
for each training iteration $t = 1, 2, \ldots$ **do**
  Set reference policy $\theta_{\text{ref}} \leftarrow \theta$.
  Sample a batch of prompts $\{\mathbf{x}_i\}_{i=1}^{B}$.
  for each prompt $\mathbf{x}_i$ **do**
    Sample K completions $\{\mathbf{y}_i^{(k)}\}_{k=1}^{K} \sim p_{\theta_{\text{ref}}}(\cdot \mid \mathbf{x}_i)$.
    Evaluate objective rewards $r_m(\mathbf{x}_i, \mathbf{y}_i^{(k)})$ and scalar scores $s_\lambda(\mathbf{x}_i, \mathbf{y}_i^{(k)})$.
    Compute group-normalized GRPO advantages $\{A_{i,k}^b\}_{k=1}^{K}$ from $\{s_\lambda(\mathbf{x}_i, \mathbf{y}_i^{(k)})\}_{k=1}^{K}$.
  end for
  Inner update (GRPO). Update θ by optimizing the standard clipped GRPO surrogate using θ_{ref} and $\{A_{i,k}^b\}$.
  COVER statistics. Using the same batch, compute completion-level clipped advantage weights $w(\mathbf{x}_i, \mathbf{y}_i^{(k)}; \theta)$.
  Compute per-objective covariance $\{c_m\}_{m=1}^{M}$ by aggregating within each prompt group and averaging across batch.
  Update EMA: $\bar{c}_m \leftarrow (1 - \tau)\bar{c}_m + \tau c_m$ for all m.
  Compute deficits $\delta_m \leftarrow [\, c_m^* - \bar{c}_m \,]_+$.
  Outer update (COVER). Update log-weights $u_m \leftarrow u_m + \eta_\lambda \, \delta_m$ and set $\lambda_m \leftarrow \exp(u_m)$ for all m.
end for

Algorithm 2 GradNorm on top of GRPO

Inputs: Objectives $\{r_m\}_{m=1}^{M}$, GradNorm exponent α, weight learning rate η_w.
Initialize weights $w_m \leftarrow 1$ for all m; initialize reference losses $L_m^{(0)} \leftarrow$ `unset`.
for each training iteration $t = 1, 2, \ldots$ **do**
  Same as COVER. Sample prompts and K completions; compute rewards and GRPO quantities needed to form a clipped surrogate.
  Per-objective gradients. Compute each objective's GRPO surrogate loss value L_m and its policy gradient g_m (using the same KL/entropy regularization as the base run).
  if $L_m^{(0)}$ is unset **then**
    Set $L_m^{(0)} \leftarrow L_m$ for all m.
  end if
  GradNorm targets. Compute gradient norms $G_m \leftarrow \|g_m\|_2$ and each objective's relative training rate from the loss ratio $L_m / L_m^{(0)}$; convert it to a target scaled norm using exponent α.
  Weight update. Update w_m with step size η_w to match the scaled norms $w_m G_m$ to their targets; renormalize weights.
  Policy update. Apply one optimization step using the weighted gradient $\sum_{m=1}^{M} w_m g_m$.
end for

Algorithm 3 Nash-MTL on top of GRPO

Inputs: Objectives $\{r_m\}_{m=1}^{M}$, Nash-MTL optimization iterations T_{Nash}, Nash learning rate η_α, numerical constant ϵ, maximum weight $\alpha_{\max}$.
Initialize Nash bargaining weights $\alpha_m \leftarrow 1$ for all m.
for each training iteration $t = 1, 2, \ldots$ **do**
  Same as COVER. Sample prompts and K completions; compute per-objective rewards, per-objective advantages, and GRPO quantities needed to form a clipped surrogate.
  Per-objective gradients. For each objective m, compute its GRPO surrogate loss value L_m and policy gradient g_m using the same KL/entropy regularization as the base run.
  Gradient matrix. Form the per-objective gradient matrix $J \in \mathbb{R}^{M \times P}$ by stacking flattened gradients, where the m-th row is $g_m^\top$.
  Nash bargaining weights. Compute the Gram matrix $G \leftarrow JJ^\top$ and approximately solve the Nash-MTL bargaining problem for $\boldsymbol{\alpha} \in \mathbb{R}_+^M$, warm-started from the previous $\boldsymbol{\alpha}$, such that $\alpha_m(G\boldsymbol{\alpha})_m \approx 1$. Clamp $\boldsymbol{\alpha}$ to $[\epsilon, \alpha_{\max}]$ for numerical stability.
  Policy update. Apply one optimization step using the Nash-weighted gradient $\sum_{m=1}^{M} \alpha_m g_m$.
end for

Algorithm 4 FAMO on top of GRPO

Inputs: Objectives $\{r_m\}_{m=1}^{M}$, weight learning rate η_w, weight decay γ, numerical constant ϵ, loss margin δ.
Initialize weight logits $\mathbf{w} \leftarrow \mathbf{0}$ and lower bounds $L_m^{\min} \leftarrow$ `unset`.
for each training iteration $t = 1, 2, \ldots$ **do**
  Same as COVER. Sample prompts and K completions; compute per-objective rewards, advantages, and GRPO quantities needed to form a clipped surrogate.
  Per-objective losses. Compute each objective's GRPO surrogate loss L_m using the same KL/entropy regularization as the base run.
  if $L_m^{\min}$ is unset **then**
    Set $L_m^{\min} \leftarrow L_m - \delta$ for all m.
  end if
  FAMO weighted loss. Let $\mathbf{z} \leftarrow \mathrm{softmax}(\mathbf{w})$, $d_m \leftarrow L_m - L_m^{\min} + \epsilon$, and $c \leftarrow \sum_m z_m/d_m$; form $L_{\text{FAMO}} \leftarrow \sum_m (z_m/c) \log d_m$.
  Policy update. Apply one optimization step using $\nabla_\theta L_{\text{FAMO}}$.
  Weight update. Recompute post-update losses $\widetilde{L}_m$ and set $\Delta_m \leftarrow \log(L_m - L_m^{\min} + \epsilon) - \log(\widetilde{L}_m - L_m^{\min} + \epsilon)$.
  Update $\mathbf{w}$ with Adam using the gradient induced by $\boldsymbol{\Delta}$ through $\mathbf{z} = \mathrm{softmax}(\mathbf{w})$.
end for

Algorithm 5 MGDA on top of GRPO

Inputs: Objectives $\{r_m\}_{m=1}^{M}$.
for each training iteration $t = 1, 2, \ldots$ **do**
  Same as COVER. Sample prompts and K completions; compute rewards and GRPO quantities needed to form a clipped surrogate.
  Per-objective gradients. Compute each objective's GRPO surrogate policy gradient g_m (using the same KL/entropy regularization as the base run).
  MGDA weights. Solve for simplex weights $w \in \Delta^M$ that minimize the norm of the combined gradient, i.e., find the minimum-norm convex combination of $\{g_m\}_{m=1}^{M}$.
  Policy update. Apply one optimization step using the weighted gradient $\sum_{m=1}^{M} w_m g_m$.
end for

Algorithm 6 Smooth Tchebycheff Scalarization with Online Mirror Descent on top of GRPO

Inputs: Objectives $\{r_m\}_{m=1}^M$, preference weights $\{\lambda_m\}_{m=1}^M$, OMD learning rate α.
for each training iteration $t = 1, 2, \ldots$ **do**
  Same as COVER. Sample prompts and K completions, compute rewards and GRPO quantities needed to form a clipped surrogate.
  Estimate objective values. For each objective, sum token-level rewards into sequence-level returns and compute the batch mean objective value R_m.
  Update the reference point. For each objective, update z_m using the running maximum sequence-level return observed so far.
  Compute smooth weights. Convert the current log-weights into smooth indicator weights by $w \leftarrow \mathrm{softmax}(u)$.
  Scalarize rewards. For each token, compute the scalar reward $s(\cdot) \leftarrow \sum_{m=1}^M w_m r_m(\cdot)$.
  Update log-weights by OMD. For each objective, update $u_m \leftarrow u_m + \alpha\lambda_m(z_m - R_m)$.
  Policy update. Compute advantages from the scalar reward s and apply one optimization step.
end for

Algorithm 7 Lagrangian Primal-Dual Method on top of GRPO

Inputs: Primary objective r_0, constraint objectives $\{r_k\}_{k=1}^K$, target constraint rewards $\{c_k\}$, dual learning rate η_λ.
Initialize Lagrange multipliers $\lambda_k \leftarrow 0$ for all k.
for each training iteration $t = 1, 2, \ldots$ **do**
  Same as COVER. Sample prompts and K completions; compute per-objective token-level rewards $\{r_0, r_1, \ldots, r_K\}$ and GRPO quantities needed to form a clipped surrogate.
  Per-objective advantages. Compute advantages for each objective separately, yielding A_0 (primary) and $\{A_k\}_{k=1}^K$ (constraints).
  Dual update. Estimate each constraint objective's average reward on the current batch, compute the reward gaps $c_k - \mathbb{E}[r_k]$, and update multipliers by dual ascent: $\lambda_k \leftarrow \max\{0,\ \lambda_k + \eta_\lambda\,(c_k - \mathbb{E}[r_k])\}$ for all k.
  Primal scalarization. Form a Lagrangian-weighted advantage: $A \leftarrow A_0 + \sum_{k=1}^K \lambda_k A_k$.
  Policy update. Apply one optimization step using the combined advantage A.
end for

F HYPERPARAMETERS

Table 11: Hyperparameter summary for different scalarization algorithms. Unless noted, all runs use the shared policy update and compute settings listed below.

Method	Hyperparameters
Lagrangian primal-dual	On MATH-500 the primary objective is accuracy, with constraints on conciseness and clarity (targets $[0.9, 0.9]$); on IFEval the primary objective is constraint following, with constraints on helpfulness (≥ 0.70) and conciseness (≥ 0.85); dual learning rate $\eta_\lambda = 0.01$; KL coefficient $= 0$.
Smooth Tchebycheff	Preference weights $\lambda_m = [0.333, 0.333, 0.334]$; OMD learning rate $\alpha = 0.3$.
MGDA	Same shared settings, except KL coefficient $= 0$.
GradNorm	Exponent $\alpha = 1.5$; weight learning rate $\eta_w = 0.025$.
Nash-MTL	Nash optimization iterations $T_{\text{Nash}} = 20$; weight learning rate $\eta_\alpha = 0.1$; numerical constant $\epsilon = 10^{-8}$; maximum weight $\alpha_{\max} = 1000$.
FAMO	Weight learning rate $\eta_w = 0.025$; weight decay $\gamma = 10^{-5}$; numerical constant $\epsilon = 10^{-8}$; loss margin $\delta = 1.0$.
COVER	Initial weights $\lambda^{(0)} = [0.333, 0.333, 0.334]$; weight learning rate $\eta_\lambda = 0.05$; covariance targets $c^* = [0.15, 0.08, 0.08]$; EMA rate $\tau = 0.1$.
Shared policy update	Learning rate $\eta = 1\text{e}^{-6}$; LR scheduler is constant; batch size $= 32$; rollouts $K = 16$; max prompt length $= 1024$; max response length $= 1024$; entropy coefficient $= 0$; KL coefficient $= 0.001$; clipping coefficient $\epsilon = 0.2$ for GRPO; epochs $= 90$; optimization backend is FSDP.
Shared compute	GPUs are $4\times$ Nvidia L40 48 GB; inference engine is vLLM.