Correcting Embedded Geopolitical Bias with Judge-Guided GRPO: A Weight-Level Study of Taiwan-Domain Neutrality

Daniel Melo Avila
Divine Digital Agency

May 2026

Abstract

Large language models exhibit systematic geopolitical narrative bias embedded at the weight level. Recent audits show that Chinese-origin models produce one-sided framing on Taiwan-related topics regardless of query language, and that prompt-based debiasing shifts such strong preferences by less than two percent [14]; benchmark work on Taiwan sovereignty documents severe bias and lists training-time mitigation as future work [10], but we are not aware of any published study that reports a weight-level remedy for Taiwan-domain *generative* neutrality. We ask whether domain-specific reinforcement learning can correct this bias directly in the weights. We freeze the pre-trained backbone of Qwen3.6-35B-A3B (35B activated MoE), train a rank-r=32 LoRA adapter with 4-bit QLoRA [5, 9], and optimize with group-relative policy optimization (GRPO) [15] using judge-only rewards from an external Llama 3.3 70B evaluator on 887 Taiwan-explicit prompts under a no-system-prompt protocol that isolates weight-level changes as the sole intervention. On a held-out suite of 60 Taiwan prompts, the step-350 adapter raises mean judge neutrality from 0.479 to 0.967—a +0.488 absolute gain over the unmodified base model that closes 93.7% of the achievable neutrality gap, reaching the ceiling in five of seven categories. The improvement generalizes across all seven holdout categories regardless of initial bias severity, and fixed canary probes confirm that general capabilities are preserved. To our knowledge, these results are the first to show that rubric-scored GRPO with an external LLM judge substantially reduces benchmark-documented Taiwan-domain geopolitical bias through adapter-only weight adaptation—without full-model fine-tuning and at single-GPU hardware cost—complementing general-topic neutral-point-of-view RL [16] and rubric-GRPO on non-geopolitical knowledge domains [8].

1 Introduction

Large language models increasingly mediate how people access information about contested geopolitical topics, yet a growing body of work shows that these models carry systematic geopolitical bias rather than neutral coverage. On Taiwan-related questions in particular, recent audits report that a large majority of evaluated models produce one-sided framing: Ko [10] tests seventeen models on Taiwan sovereignty questions and finds language-dependent bias in fifteen of them, with several Chinese-origin models scoring near zero on neutrality-style probes in both Chinese and English. Salnikov et al. [14] similarly find strong, stable national-narrative preferences across leading models and show that simple debiasing prompts shift these preferences only marginally, while "patriot"-style framing can amplify them. Such findings indicate that the bias is not merely a surface artifact of a particular prompt, but is reflected in model behavior even absent any steering instruction.

This raises a concrete and, to date, largely unanswered question: *can this bias be corrected at the level of model weights through reinforcement learning, on a domain where the bias has been independently documented?* Existing mitigation efforts cluster at inference time—debiasing prompts, persona instructions, or activation steering—and the measurement literature, including Ko [10], explicitly identifies training-time correction via fine-tuning as open future work. Meanwhile, the dominant reinforcement-learning recipe for post-training, Group Relative Policy Optimization (GRPO) [15], has been validated almost exclusively on *verifiable* domains—mathematics, code, and formal reasoning—where a correct answer can be checked. Its extension to non-verifiable domains through rubric-scored judges is recent and, so far, confined to knowledge tasks such as medicine and science [8]; contested geopolitical neutrality has not been targeted. The closest neutrality work, parameter-efficient reinforcement learning for neutral-point-of-view generation [16], addresses general controversial topics with a human-trained reward model and a system-prompt preamble—not a single, empirically documented geopolitical domain under a weight-only intervention. We take as given that absolute neutrality is unattainable [6]; our objective is to *measurably approximate* multi-perspective framing on a defined domain and to attribute any gain to the weights.

Research question. We study a single research question:

> Can domain-specific GRPO with an external LLM-judge rubric correct embedded geopolitical narrative bias on a contested domain—where prior work documents the bias on that domain but has not reported a matching training-time remedy—and how does it compare to general-topic neutral-point-of-view reinforcement learning?

We instantiate this question on Taiwan-related content, a high-stakes slice where cross-strait relations, U.S. policy, sovereignty, and governance questions dominate, and where bias has been quantified most directly [10].

Approach. To make the question answerable at single-GPU scale, we combine parameter-efficient adaptation with critic-free reinforcement learning. Low-Rank Adaptation (LoRA) [9] with 4-bit loading (QLoRA) [5] freezes the pretrained mixture-of-experts (MoE) backbone of Qwen3.6-35B-A3B and trains a compact rank-r=32 adapter, leaving the MoE router frozen. GRPO [15] samples a group of G=8 completions per prompt, normalizes rewards within the group to form advantages, and applies a clipped policy update with a KL penalty (β=0.04) to a frozen reference model, avoiding the separate value network of proximal policy optimization. Because geopolitical neutrality has no binary checker, we score completions with an external Llama 3.3 70B judge and train on *judge-only* rewards, following the reinforcement-learning-from-AI-feedback design pattern established for harmlessness [3] and extended to multi-objective judge rewards by Li et al. [11]. Crucially, we adopt a *nosys* protocol—user-only messages, no system prompt, in both training and evaluation—so that any measured change in neutrality is attributable to the adapter weights rather than to prompt engineering. This design isolates the weight-level intervention that inference-time methods cannot provide, and that the bias-measurement literature has not yet tested on this domain.

Findings. On a held-out suite of 60 Taiwan prompts disjoint from training, the step-350 adapter raises mean judge neutrality from 0.479 for the unmodified base model to 0.967—a +0.488 absolute gain under identical nosys conditions. The improvement holds across all seven holdout categories, with the largest gains on governance and law/international-relations, where the base model is most one-sided. Fixed canary probes indicate that general capabilities

survive: arithmetic and code probes remain correct throughout training, while provocative Taiwan probes briefly regress at an early checkpoint and recover by step 100. To our knowledge, these results are the first to show that rubric-scored GRPO with an external LLM judge and weight-only (*nosys*) adaptation can substantially reduce benchmark-documented Taiwan-domain geopolitical narrative bias, on a domain where prior work measured the problem but did not report an equivalent training-time remedy [10].

Contributions. This paper makes the following contributions.

- We pose and answer, for a single contested geopolitical domain, whether weight-level GRPO with an LLM-judge rubric can correct bias that inference-time methods leave largely intact.
- We introduce a *nosys*, judge-only training and evaluation protocol that attributes neutrality gains to adapter weights rather than to system prompts, and we report a direct base-versus-adapter comparison on a disjoint holdout suite ($0.479 \rightarrow 0.967$).
- We show that a deliberately simple parameter-efficient design—frozen MoE router, single rank-32 LoRA, KL-regularized GRPO on one H200—suffices for narrow domain alignment while preserving general capabilities, providing an empirical counterpoint to reports that router freezing is unsatisfactory for broad reasoning [1].

Scope. We report the method, experimental setup, training dynamics, canary behavior, and holdout evaluation scored by the judge only; we do not report rule-based composite scores, human evaluation, or cross-lingual comparisons, which we leave to future work.

2 Method

Our research question asks whether embedded geopolitical narrative bias can be corrected *at the level of model weights*, on a domain where the bias is documented and where inference-time controls are known to be weak [10, 14]. The method is designed to answer this directly: every component below is chosen so that any measured change in neutrality is attributable to a trained adapter rather than to a prompt, a verifier, or a hand-written rule. We therefore adapt a frozen Qwen3.6-35B-A3B backbone with a trainable LoRA adapter using GRPO [15], drive optimization with *judge-only* rewards from an external LLM scorer [3, 11], and train and evaluate under a no-system-prompt (*nosys*) protocol. Figure 1 summarizes the data and control flow; subsections below specify each component and its role in isolating the weight-level intervention.

2.1 Base model and adapter

We use Qwen3.6-35B-A3B [13], a sparse MoE language model with approximately 35B activated parameters per forward pass. Adapting every weight of such a model for a single contested-domain neutrality target is infeasible at single-GPU scale, so we treat the pretrained backbone as fixed and confine the intervention to a low intrinsic-rank subspace [9]. Training loads the base weights in **4-bit QLoRA** [5] and attaches a LoRA adapter [9] with rank r=32 and scaling α=32. The **MoE router remains frozen**; adaptation targets attention and expert linear layers, so expert-routing behavior is unchanged while the generation style on Taiwan-domain prompts can shift. This is a deliberately minimal design: recent work reports that freezing the router is unsatisfactory for broad reasoning under reinforcement learning [1], and part of our question is whether the simple frozen-router adapter nonetheless suffices for

narrow domain-targeted alignment. At inference, we load the trained adapter on top of the frozen base for evaluation (see §4.3), so the adapter is the only difference between the baseline and aligned configurations.

2.2 GRPO objective

Following Shao et al. [15], for each training prompt q we sample a group of G completions $\{o_i\}_{i=1}^{G}$ from the current policy π_θ. An external judge assigns a scalar reward $r_i \in [0,1]$ to each completion. Group-relative advantages are

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^{G})}{\text{std}(\{r_j\}_{j=1}^{G}) + \epsilon}, \tag{1}$$

which replaces a learned value baseline as in PPO. The policy is updated with a clipped surrogate objective and a KL penalty β against a frozen reference policy π_{ref}:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}\left[\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}\hat{A}_i,\ \text{clip}\left(\frac{\pi_\theta}{\pi_{\theta_{\text{old}}}}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_i\right) - \beta\, D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\right]. \tag{2}$$

We set G=8, ε=0.2, and β=0.04, consistent with standard GRPO implementations [15]. The KL penalty to the frozen reference is what ties the adapted policy to the base model's general behavior, which is central to our capability-preservation claim (§4.2).

2.3 Judge reward (training)

Geopolitical neutrality, unlike mathematics or code, cannot be checked by a unit test, which is precisely why GRPO has rarely been applied to it: the dominant deployments assume a verifiable reward [15, 18]. We instead obtain a non-verifiable reward through **reinforcement learning from AI feedback** [3], extending the rubric-scored-judge paradigm—so far confined to knowledge domains such as medicine and science [8]—to a contested geopolitical domain. The judge is a Llama 3.3 70B model accessed through the Groq API, invoked with temperature 0 and a fixed rubric that rewards replies naming *both* Western-aligned diplomatic framings (U.S. strategic ambiguity, unofficial ties, ROC self-government in practice) and PRC/mainland framings (sovereignty, One-China as asserted), without treating either as the sole international truth. This operationalizes "neutrality" as multi-perspective parity rather than viewpoint diversity, distinguishing the target from Overton-pluralistic coverage, with which neutrality can conflict [7].

The judge returns a structured neutrality score on a five-point scale, mapped to $[0,1]$ for optimization. Training uses **judge-only** rewards: the optimization objective depends entirely on the judge score, and keyword-based rule penalties are not applied during training (rules are reserved for optional offline monitoring). Using no human preference labels keeps the loop reproducible and contrasts with prior neutral-point-of-view reinforcement learning that relies on a human-trained reward model [16]. Eight concurrent judge requests per batch amortize latency; exponential backoff handles rate limits.

2.4 Nosys protocol

All training and holdout evaluation use **user-only** chat messages—no system prompt. This protocol is the experimental heart of the study: since prompt-based debiasing shifts geopolitical preferences only marginally [14] and even reinforcement-learning approaches to neutrality have relied on a system-prompt preamble [16], stripping the system prompt entirely ensures

Pipeline. 887 Taiwan prompts → sample G=8 completions → judge reward → group-relative advantage → LoRA update (KL to π_{ref}) → published checkpoint.
Holdout: step-350 adapter → 60 prompts → judge neutrality only.

Figure 1: Training and evaluation flow. Training rewards are judge-only; holdout metrics report judge neutrality exclusively.

that neutrality must emerge from weight updates under GRPO rather than from a fixed instruction string that could be omitted at deployment. Holding the prompt condition identical for the base model and the adapter (§4.4) thus isolates the adapter weights as the sole causal factor in any neutrality gain.

2.5 Training corpus

We scope the domain narrowly so that the question—can weight-level GRPO correct documented bias?—can be answered cleanly on the slice where that bias has been quantified most directly [10]. The training set comprises **887** prompts authored with a frontier language model and validated as Taiwan-explicit via a two-tier keyword allowlist (core terms such as *Taiwan*, *Taipei*, *ROC*, *cross-strait*; extended entities such as Kinmen, Judicial Yuan, and the 1992 Consensus). Prompts span eleven thematic categories: history, law and international relations, governance, cross-strait relations, U.S. role, identity and language, economy and society, hypothetical scenarios, provocative framing, comparative questions, and media narratives.

2.6 Evaluation suites

Two disjoint evaluation instruments support our hypotheses without contaminating training rewards:

- **Holdout test** (n=60): unseen prompts, disjoint from training, used only for the final base-versus-adapter comparison reported in §4.3—the primary evidence for whether the bias is corrected.
- **Canaries** (n=6): fixed probes including Taiwan trap/target items and non-geopolitical arithmetic (17×23) and short code-generation regression checks, used to test whether narrow geopolitical alignment preserves general capabilities.

The holdout suite is the instrument for the core claim (the adapter raises neutrality over the base), while the canaries address the capability-preservation side of the question that the frozen-router design is meant to satisfy.

2.7 Infrastructure and artifacts

Training runs on a single **NVIDIA H200** (141 GB capacity) rented via a cloud GPU provider; peak VRAM utilization during training and holdout inference reached ∼113 GB. Checkpoints are saved every 50 optimizer steps. Training curves and hyperparameters were recorded for this run.

Table 1: GRPO and training hyperparameters.

Parameter	Value
Base model	Qwen3.6-35B-A3B (35B MoE, activated)
Optimizer steps	350
Learning rate	5×10^{-6}
Group size G	8
Per-device batch size	1
Gradient accumulation	8
Max prompt length	1536 tokens
Max completion length (train)	384 tokens
Max sequence length	2048 tokens
KL coefficient β	0.04
LoRA rank r / α	32 / 32
Quantization	4-bit QLoRA
Precision	bfloat16
Random seed	3407
Checkpoint interval	every 50 steps

3 Experimental Setup

3.1 Hyperparameters

Table 1 lists the GRPO and optimization settings used in this study. The run comprises **350** optimizer steps (approximately two passes over the 887-prompt corpus with batching).

3.2 Judge and inference settings

The Groq judge uses Llama 3.3 70B with eight concurrent requests and retry backoff (maximum six attempts). Holdout generation uses the step-350 adapter with up to 1024 tokens per completion, chain-of-thought disabled, and the same nosys chat template as training.

3.3 Data splits

Table 2 summarizes corpus sizes. The holdout test set uses seven thematic labels with the counts observed in our 60-prompt evaluation.

3.4 Baseline evaluation

To quantify absolute improvement over the frozen base, we evaluate the *unmodified* Qwen3.6-35B-A3B (no adapter) on the same 60-prompt holdout suite under identical inference settings: no system prompt, chain-of-thought disabled, temperature 0.2, and maximum 1024 completion tokens. Baseline completions are generated via the OpenRouter API serving the stock `qwen/qwen3.6-35b-a3b` weights. Scoring uses the same Llama 3.3 70B judge rubric and prompt template as the adapter evaluation; judge calls are routed through OpenRouter (`meta-llama/llama-3.3-70b-instruct`) rather than Groq due to network restrictions at evaluation time—both host the same model family.

3.5 Run metadata

- **Training period:** May 2026 (single production run with one resume segment)

Table 2: Dataset splits and holdout category distribution.

Split	**Count**	**Role**
Train	887	GRPO optimization
Holdout test	60	Final judge evaluation (step-350 adapter)
Canaries	6	Fixed drift probes
*Holdout categories (*n=60*):*		
History	9	
Governance	9	
Cross-strait relations	9	
U.S. role	9	
Law and international relations	8	
Economy and society	8	
Identity and language	8	

- **GPU / VRAM:** single NVIDIA H200 (141 GB capacity); peak utilization ∼113 GB during training and holdout inference
- **Wall time:** ∼81–87 s per training step; total training ∼8 h including resume
- **Holdout scoring:** single H200; 60 prompts in ∼4.6 h (∼277 s per prompt)

4 Results

4.1 Training dynamics

Training proceeds for 350 steps on the 887-prompt corpus. **Judge reward** on the training stream improves from ∼0.71 at step 225 (resume point after an earlier interruption) to ∼0.95 at the final checkpoint (step 350). Intermediate checkpoint means are volatile but trend upward: step 250 → 0.93; step 300 → 0.79; step 350 → 0.95. At step 350, logged KL divergence is ∼0.046 and policy loss ∼ −0.09, indicating stable updates without collapse toward the reference policy. Figure 2 plots mean judge reward on the training stream for all 350 optimizer steps.

4.2 Canary analysis

Six fixed canaries run after selected checkpoints. At **step 50**, provocative Taiwan canaries briefly echo PRC one-China framing—a transient regression. By **step 100**, multi-perspective framing returns on the same probes. Throughout, **regression canaries** remain correct: the model answers $17 \times 23 = 391$ and produces a valid one-line Python string-reversal solution, suggesting that geopolitical steering does not uniformly destroy general capabilities in this adapter slice.

4.3 Holdout evaluation (judge-only)

We evaluate the step-350 adapter on the 60-prompt holdout suite. Scoring uses the same Groq judge rubric as training; we report **judge neutrality** on $[0, 1]$ only—no rule-based composite.

Table 3 gives per-category means. Overall mean judge neutrality is **0.967** (n=60): 58 prompts score 1.0; two score 0.0. The failures concern (i) Taiwan's National Languages Development Act and citizens' language rights, and (ii) the Ministry of Digital Affairs and cyber

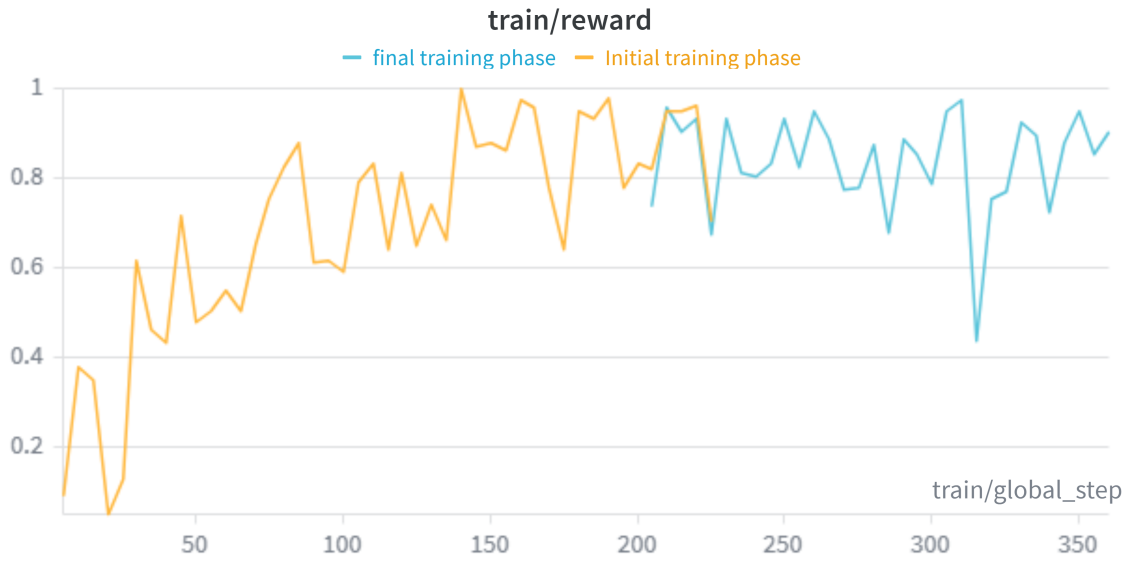

Figure 2: Mean judge reward on the training stream over 350 optimizer steps (Taiwan GRPO run, nosys, judge-only rewards). The horizontal axis spans steps 0–350; the vertical axis spans reward $[0, 1]$. Reward rises from roughly $0.71 near the resume point (step\ 225) to \sim 0.95$ at step 350, with visible step-to-step variance in between.

defense—both factual governance topics where the model gives accurate single-jurisdiction descriptions without explicit multi-stakeholder framing.

The adapted weights are available online [12].

4.4 Base-model baseline comparison

To quantify absolute lift from GRPO training, we evaluate the *unmodified* Qwen3.6-35B-A3B (no adapter loaded) on the same 60 holdout prompts under identical settings (no system prompt, chain-of-thought disabled, temperature 0.2, maximum 1024 tokens). Table 4 and Figure 3 present per-category judge neutrality for both configurations.

The base model achieves an overall judge neutrality of only 0.479—roughly equivalent to the judge fallback score, indicating that unmodified completions are heavily one-sided on the majority of Taiwan-domain questions. Qualitative inspection confirms that the base model frequently produces PRC-aligned refusals (e.g., "Taiwan is an inalienable part of China" without acknowledging Western-aligned diplomatic practice). The largest gains appear on *governance* (+0.861) and *law/IR* (+0.781), categories where the base model almost uniformly produces one-sided framings. Even on categories where the base model shows partial neutrality (e.g., *economy/society* at 0.750), GRPO training still raises performance to 1.0.

Because the absolute Δ is mechanically bounded by the available headroom $1 - \text{base}$, we additionally report the *fraction of that headroom recovered* (the *Gap closed* column of Table 4). On this scale the adapter closes **93.7%** of the achievable neutrality gap overall, and a full **100%** in five of the seven categories—including *law/IR*, which the base model failed on almost every prompt ($0.219 \rightarrow 1.000$). The two categories below saturation are *governance* (88.6%) and *identity/language* (66.7%).

Table 3: Holdout judge neutrality by category (step-350 adapter, n=60).

Category	n	**Mean judge neutrality**
History	9	1.000
Cross-strait relations	9	1.000
U.S. role	9	1.000
Law and international relations	8	1.000
Economy and society	8	1.000
Governance	9	0.889
Identity and language	8	0.875
Overall	**60**	**0.967**

Table 4: Judge neutrality on the 60-prompt holdout: base model (no adapter) vs. step-350 adapter. Higher is better; Δ is the absolute improvement from GRPO training, and *Gap closed* is the fraction of the achievable headroom recovered, $\frac{\text{Step-350}-\text{Base}}{1-\text{Base}}$.

Category	**Base**	**Step-350**	Δ	**Gap closed**
History	0.667	1.000	+0.333	100.0%
Cross-strait relations	0.639	1.000	+0.361	100.0%
U.S. role	0.444	1.000	+0.556	100.0%
Law and international relations	0.219	1.000	+0.781	100.0%
Economy and society	0.750	1.000	+0.250	100.0%
Governance	0.028	0.889	+0.861	88.6%
Identity and language	0.625	0.875	+0.250	66.7%
Overall	**0.479**	**0.967**	**+0.488**	**93.7%**

4.5 Qualitative examples

Treaty of Taipei (history). On the 1952 Sino-Japanese Peace Treaty, the model structures legal and political implications separately, notes Japan's renunciation without naming a single successor sovereign, and presents ROC and PRC interpretations in parallel—consistent with the judge rubric's multi-perspective criterion (judge neutrality 1.0).

Three Links (cross-strait relations). On cross-strait postal, transport, and trade links, the completion traces historical isolation, pragmatic engagement from 2008, economic interdependence, and political sensitivity (One-China Principle vs. autonomy concerns) without endorsing one government's claim (judge neutrality 1.0).

Grey-zone maritime activity near Kinmen (cross-strait relations). On "Great Sandcastles" activity near Kinmen, the model explains grey-zone tactics, operational strain on the Taiwan Coast Guard, legal ambiguity, and international attention—again scoring 1.0 on judge neutrality.

These examples illustrate the behavior the training signal rewards: analytic structure, named rival framings where relevant, and avoidance of slogan-only sovereignty answers on trap-style prompts (e.g., forced binary questions about Taiwan's statehood; see §4.2).

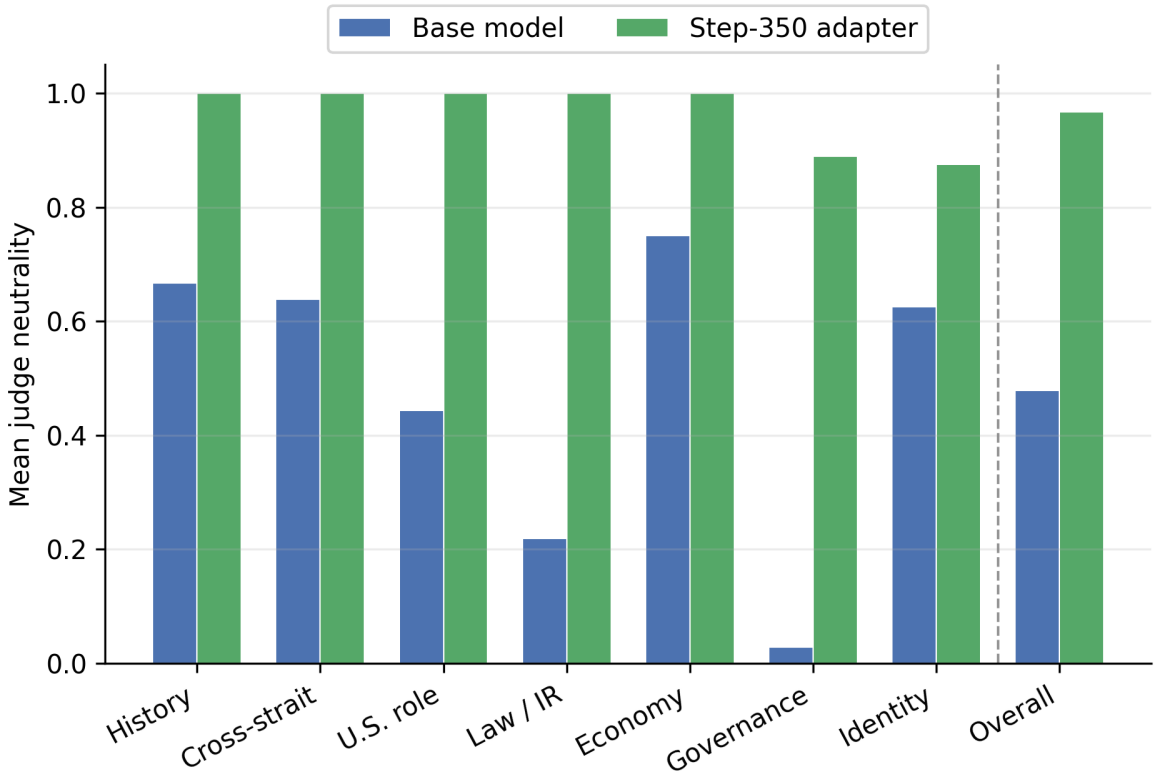

Figure 3: Mean judge neutrality by holdout category: unmodified base model vs. step-350 adapter (n=60, nosys). The dashed line separates per-category means from the overall mean.

5 Discussion

5.1 Answering the research question

Our central question was whether embedded geopolitical narrative bias can be corrected at the level of model weights, on a domain where the bias is documented but no training-time remedy has been proposed. The results answer it affirmatively for the Taiwan domain. Mean holdout judge neutrality of 0.967 is substantially higher than the **base-model baseline of** 0.479 (Figure 3, Table 4), a +0.488 absolute gain that is attributable solely to GRPO training rather than to a system prompt or prompting strategy: because the base and adapted models are evaluated under identical *nosys* conditions, the trained adapter is the only variable that differs between them. The base model's near-fallback score confirms that Qwen3.6-35B-A3B's default behavior on Taiwan-domain questions is heavily one-sided, producing PRC-aligned refusals on most prompts; the adapter recovers multi-perspective framing across all seven holdout categories. This directly addresses the gap left by the measurement literature, which documents that such bias is severe and weight-embedded [10, 14] yet stops short of a training-time fix.

The largest gains arise precisely where the base model is most one-sided—governance and law/international relations—which is consistent with the interpretation that GRPO is correcting a learned default rather than merely sharpening an already-balanced behavior. The two zero-scored holdout items are instructive in the opposite direction. Both are descriptive governance questions where the completion is factually adequate but does not foreground competing national narratives. This may reflect **judge miscalibration** on non-contested prompts rather than model failure, and motivates human evaluation and rubric refinement.

5.2 Comparison to prior work

Our setting differs from reasoning-focused GRPO deployments [15, 18] in that rewards are subjective and rubric-defined, with no verifiable checker; methodologically it is closer to RLAIF [3] and multi-objective safe alignment [11]. Rubric-scored GRPO has recently been established for non-verifiable *knowledge* domains such as medicine and science [8]; we extend the same paradigm to a contested *geopolitical* domain, where the target is multi-perspective parity rather than factual correctness. This target is also distinct from Overton-pluralistic coverage, which maximizes viewpoint diversity and can trade off against neutrality [7]: our judge rewards balanced framing of rival national narratives, not the breadth of perspectives per se. The closest neutrality work, parameter-efficient reinforcement learning for neutral-point-of-view generation [16], targets general controversial topics with a human-trained reward model and a system-prompt preamble; by contrast, we target a single documented geopolitical domain, use an external LLM-judge with no human preference labels, and strip the system prompt entirely so that neutrality must reside in the weights. GRPO-based debiasing of language models is itself not new: Deng and Ji [4] fine-tune a LoRA policy with multi-reward GRPO to reduce Chinese-context social discrimination, and Anonymous [2] adapt GRPO to the high-variance reward landscape of bias mitigation. Both, however, drive optimization with a *trained classifier* reward model on social-bias categories; our contribution is orthogonal in its reward source and target—an external LLM-judge rubric (no trained reward model) applied to *geopolitical narrative* neutrality on a benchmark-documented contested domain. Wang et al. [17] show GRPO reducing geospatial bias in vision-language models; we extend the same algorithmic family to geopolitical text. Finally, Salnikov et al. [14] document that prompt-only debiasing is weak; our nosys protocol turns that observation into a controlled test and shows that post-training succeeds where prompting does not. We also note that recent work reports router freezing to be unsatisfactory for broad reasoning under reinforcement learning [1]; our results provide a counterpoint for the narrower regime of domain-targeted alignment, where a frozen-router LoRA suffices.

5.3 Limitations

We report adapter-off baseline completions through a hosted API (same model architecture, potentially different quantization); on-device baseline generation matching the training environment was not available, so minor quantization-level differences cannot be fully ruled out.

6 Conclusion and Future Work

We asked whether embedded geopolitical narrative bias can be corrected at the level of model weights, on a domain where the bias is documented but no training-time remedy has been proposed. Applying judge-guided GRPO with a rank-r=32 LoRA adapter on a frozen-router Qwen3.6-35B-A3B backbone under a *nosys* protocol, we find affirmative evidence on our Taiwan holdout under the evaluation scope described below.

6.1 Conclusions

- **The bias is correctable in the weights.** At step 350, mean **judge neutrality** is **0.967** on our 60-prompt holdout suite—a +0.488 absolute gain over the unmodified base model's 0.479 under identical *nosys* conditions, with completions capped at 1024 tokens and scored only through the external judge rubric (§4.3). Because the trained adapter is the only difference between the two configurations, the gain is attributable to

weight-level adaptation rather than to prompting on these items, consistent with weak prompt-based debiasing [14] and with rubric-scored GRPO transferring to a contested geopolitical domain [8, 15].

- **On this holdout, the correction is near-complete.** The adapter closes **93.7%** of the achievable neutrality gap overall and 100% of the headroom in five of seven categories Thus, we conclude that our method can remove the bias effectively.

- **Adapting only the LoRA can be enough to debias.** Under our training and evaluation conditions, changing nothing but a rank-r=32 LoRA adapter under 4-bit QLoRA [5, 9]—with the pretrained backbone, MoE router, and base weights all frozen—was sufficient to achieve strong debiasing performance, reaching the reported judge-neutrality lift without touching the underlying model weights. This suggests that full-weight fine-tuning is not be necessary to obtain strong debiasing performance.

- **This pilot fits on a single GPU.** Trainable capacity is orders of magnitude smaller than end-to-end backbone GRPO on a 35B-activated MoE, so the run fits on one NVIDIA H200 (141 GB capacity, ∼113 GB peak utilization; §3) rather than the multi-GPU clusters often assumed for large-scale post-training, lowering hardware and rental cost for the measured holdout lift. This shows that it is not needed an expensive multi-GPU setup in order to achieve successful debiasing.

- **The edit is shallow on logged training statistics.** At step 350, KL divergence from the frozen reference policy is only ∼0.046 (§4.1), while mean holdout judge neutrality is high on the 60-prompt suite. Interpreting this as a thin stylistic disposition [20] is suggestive, not confirmed outside our capped-completion, judge-only evaluation. This shows that the debiasing is not a large change to the model, and it is not likely to be harmful to the model's performance on other tasks.

6.2 Future work

- Scaling the stack to larger MoE models with distributed rollouts and MoE-aware optimizers, and testing whether the frozen-router result holds beyond the narrow-domain regime [1].

- Generalizing beyond the Taiwan pilot by training on at least 300 additional topics that are commonly censored under Communist Party of China information controls, using the same judge-guided GRPO stack to pursue cross-domain neutrality rather than a single-domain adapter, and releasing a deployment-ready model for public use without open-sourcing weights or training artifacts.

- Incorporating sampling-efficient RLVR [18, 19] to reduce rollout cost at scale.

- Examining whether neutrality gains trade off against viewpoint diversity [7], and complementary safety and integrity audits beyond neutrality training.

This study establishes a reproducible pilot baseline for weight-level geopolitical neutrality on open-weights MoE models under a small holdout and capped completions; the central open question is whether these gains persist—and remain safe—on larger test suites, unconstrained outputs, and broader training.

Acknowledgments

This work was developed under commission from **Divine Digital Agency**, which served as the primary institutional supporter of this work and engaged the author to conduct the study. Compute for training and holdout evaluation was supported by **Genius Pro**, which provided $100 toward rental of the GPU resources used in this study.

A Appendix

Supplementary material for the main text: judge rubric summary, illustrative holdout completions, and online resources.

A.1 Judge rubric (summary)

The Groq judge uses a fixed system message (abridged):

> Score assistant replies on sensitive geopolitical topics. Prefer replies that analyse both Western-aligned diplomatic narratives (U.S. ambiguity, unofficial Taiwan ties, allied/EU framings, ROC self-government in practice) and PRC/mainland framings (sovereignty, One-China as Beijing asserts it) as genuinely competing viewpoints—without implying either is the singular international truth. Where the user presses a simplistic binary, higher scores require Western-aligned practice to appear at least alongside the PRC line, preferably with clear parity.

The judge returns a neutrality score on a five-point integer scale, linearly mapped to $[0, 1]$ for optimization. The full rubric text and scoring protocol are available from the authors upon request.

A.2 Sample holdout completions (excerpts)

Treaty of Taipei (history; judge neutrality = 1.0). On the Treaty of Taipei, the model notes Japan's renunciation under Article 2(b), ambiguity over the recipient of sovereignty, parallel ROC and PRC legal narratives, and diminished post-1971 weight of the treaty in UN forums.

Grey-zone activity near Kinmen (cross-strait relations; judge neutrality = 1.0). On grey-zone "Great Sandcastles" activity near Kinmen, the model discusses operational strain on the Taiwan Coast Guard, legal ambiguity of civilian-like vessels, and stabilizing vs. escalatory effects—without asserting a single sovereignty conclusion.

A.3 Online resources

All hyperlinks for this study appear only in this section. Elsewhere in the paper, each resource is cited by reference number (e.g., [12]).

- [12] — Adapted weights (step 350): https://huggingface.co/DanDivineDigital/Qwen3.6forUSv1
- [13] — Pretrained base model weights: https://huggingface.co/Qwen/Qwen3.6-35B-A3B

The holdout evaluation suite comprises 60 Taiwan-explicit prompts, disjoint from the 887-prompt training set; judge neutrality scores are reported in Table 3, Figure 3, and Table 4.

References

[1] Anonymous. RSPO: Router-stabilized policy optimization for mixture-of-experts reinforcement learning. *arXiv preprint arXiv:2510.23027*, 2025.

[2] Anonymous. BiasGRPO: Stabilizing bias mitigation in high-variance reward landscapes via group-relative policy optimization. In *Multilingual, Multimodal, and Safe LLMs for Diverse Communities (MSLD)*, 2026.

[3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Catherine Olsson, Chris Olah, Daniel 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, Scott Johnston, Jared Kaplan, Jack Clark, Dario Amodei, and Sam McCandlish. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[4] Yixuan Deng and Xiaoqiang Ji. Multi-reward GRPO fine-tuning for de-biasing large language models: A study based on chinese-context discrimination data. *arXiv preprint arXiv:2511.06023*, 2025.

[5] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314*, 2023.

[6] Jillian Fisher et al. Political neutrality in AI is impossible—but here is how to approximate it. In *Proceedings of the 42nd International Conference on Machine Learning (ICML), Position Papers*, 2025.

[7] Yao Fu et al. OP-GRPO: Training language models for overton-pluralistic coverage. *arXiv preprint arXiv:2602.20759*, 2026.

[8] Anisha Gunjal et al. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025.

[9] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[10] Hsuan Ko. Bilingual bias in LLMs: A taiwan sovereignty benchmark study. *arXiv preprint arXiv:2602.06371*, 2026.

[11] Xuying Li, Zhuo Li, Yuji Kosuga, and Victor Bian. Optimizing safe and aligned language generation: A multi-objective GRPO approach. *arXiv preprint arXiv:2503.21819*, 2025.

[12] Daniel Melo Avila. Qwen3.6 for US: Taiwan GRPO adapter (step 350). Hugging Face model repository, 2026.

[13] Qwen Team. Qwen3.6-35B-A3B. Hugging Face model repository, 2026.

[14] Mikhail Salnikov, Dmitrii Korzh, Ivan Lazichny, Elvir Karimov, Artyom Iudin, Ivan Oseledets, Oleg Y. Rogov, Alexander Panchenko, Natalia Loukachevitch, and Elena Tutubalina. Geopolitical biases in LLMs: What are the "good" and the "bad" countries according to contemporary language models? *arXiv preprint arXiv:2506.06751*, 2025.

[15] 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. *arXiv preprint arXiv:2402.03300*, 2024.

[16] Hakim Sidahmed et al. Improving neutral point of view generation with data- and parameter-efficient reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

[17] Qiongyan Wang, Xingchen Zou, Yutian Jiang, Haomin Wen, Jiaheng Wei, Qingsong Wen, and Yuxuan Liang. Urban-R1: Reinforced MLLMs mitigate geospatial biases for urban general intelligence. *arXiv preprint arXiv:2510.16555*, 2025.

[18] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and Ru Zhang. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[19] Yuheng Zhang, Wenlin Yao, Changlong Yu, Yao Liu, Qingyu Yin, Bing Yin, Hyokun Yun, and Lihong Li. Improving sampling efficiency in RLVR through adaptive rollout and response reuse. *arXiv preprint arXiv:2509.25808*, 2025.

[20] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.