Small-Scale Outcome-Only GRPO Barely Moves the Answer Distribution

A Global-Affine Decomposition and a Paired Null, with No Accuracy Gain to Recover

Ephraiem Sarabamoun

Abstract

If outcome-only reinforcement learning mostly elicits what the base model already computes [@zhou2023lima; @lin2023unlocking], its effect on the answer distribution might be reproducible by a single global affine transform of the base logits, a scalar temperature plus a per-option bias, which is exactly temperature scaling extended with a class offset [@guo2017calibration]. We test this on Qwen2.5-1.5B-Instruct before and after outcome-only GRPO [@shao2024deepseekmath] on MMLU [@hendrycks2021mmlu], extracting the four-option answer logits of both models and fitting a global affine map from base to RL logits. Two things hold. First, GRPO barely changed the model: it altered the predicted answer on only 27 of 1000 items (2.7 percent), with no significant direction (McNemar chi-squared 0.45, p=0.50), and produced no significant accuracy gain (base 0.3780, 95 percent CI 0.3490 to 0.4090; RL 0.3820, 0.3520 to 0.4130; gain +0.0040), so the scalar elicitation fraction we set out to report is ill-defined, a ratio with a near-zero denominator. Second, the small change GRPO did make to the logits is almost entirely global: the base logits already explain 98.5 percent of the RL logit variance with no transform, a per-option bias alone raises that to 99.3 percent, and the full affine (temperature 0.989) adds essentially nothing more, the change being a near-constant upward shift on the option logits, largest on option A. For reference a global affine transform of base logits fit toward the gold answer reaches 0.3890 and a full linear map 0.4040, but both are supervised on the evaluation labels and so are recalibration ceilings, not fair comparisons to unsupervised GRPO. At this scale, then, the honest summary is that outcome-only GRPO barely moved the answer distribution, and what little it moved is a global per-option bias shift rather than an input-dependent change; there is essentially no accuracy gain for any transform to recover.

1. Introduction

The elicitation view of post-training holds that reinforcement learning and instruction tuning mostly surface capabilities the base model already has [@zhou2023lima]. A related line shows base models can be unlocked toward aligned behavior by in-context means alone, with no weight update [@lin2023unlocking]. If that is right at the level of the output distribution, then the change RL makes to a model's answer logits should be low-complexity and, in the simplest case, a global recalibration: a single temperature and a per-class bias, the same family used to calibrate classifiers [@guo2017calibration]. Outcome-only RL, which rewards only the final answer [@ouyang2022instructgpt], whether through helpful-and-harmless RLHF [@bai2022hh] or verifiable-reward RL [@shao2024deepseekmath], is the cleanest case to test, because it cannot directly shape the reasoning and might act largely by reweighting the answer distribution, much as preference optimization can be written as a closed-form reweighting of the base policy [@rafailov2023dpo]. We ask, concretely, whether GRPO's effect on the four-option MMLU answer logits is recovered by a single global affine transform of the base logits, and we report a scalar elicitation fraction, the share of GRPO's accuracy gain that such a transform reproduces.

2. Method

We reuse a base model and an outcome-only-RL model from a sibling study: Qwen2.5-1.5B-Instruct [@qwen2024report] at GRPO step zero (the base) and after 120 GRPO steps [@shao2024deepseekmath] with a correctness-only verifiable-answer reward of the style used for math word problems [@cobbe2021gsm8k]. On 1000 reserved MMLU eval items [@hendrycks2021mmlu], we prefill the assistant turn with the string that begins the answer and read the logits over the four option-letter tokens, the model's direct answer distribution [@wei2022cot], an ability present even in the zero-shot setting [@kojima2022zeroshot]. We then fit a global affine map, scalar temperature a and per-option bias vector b, from base option logits to RL

option logits by least squares and report the variance explained, and we compare it against two baselines, the identity (no transform) and a per-option bias with temperature fixed at one. We also fit a global affine transform of the base logits toward the gold answer by minimizing cross-entropy in five-fold cross-validation, the best a global recalibration can do toward correctness, and a full unconstrained linear map of the four logits as a non-global ceiling. The scalar elicitation fraction is the gold-fit affine accuracy gain over base divided by GRPO's accuracy gain over base [@guo2017calibration]. Base models carry calibrated answer knowledge [@kadavath2022know] and recoverable latent structure [@burns2022ccs], so a recalibration is a plausible mechanism; whether outcome reward can do more than recalibrate is the process-versus-outcome question [@lightman2023letsverify]. All accuracies carry percentile bootstrap 95 percent confidence intervals at seed 42, and a recompute script rebuilds them from per-example predictions.

3. Results

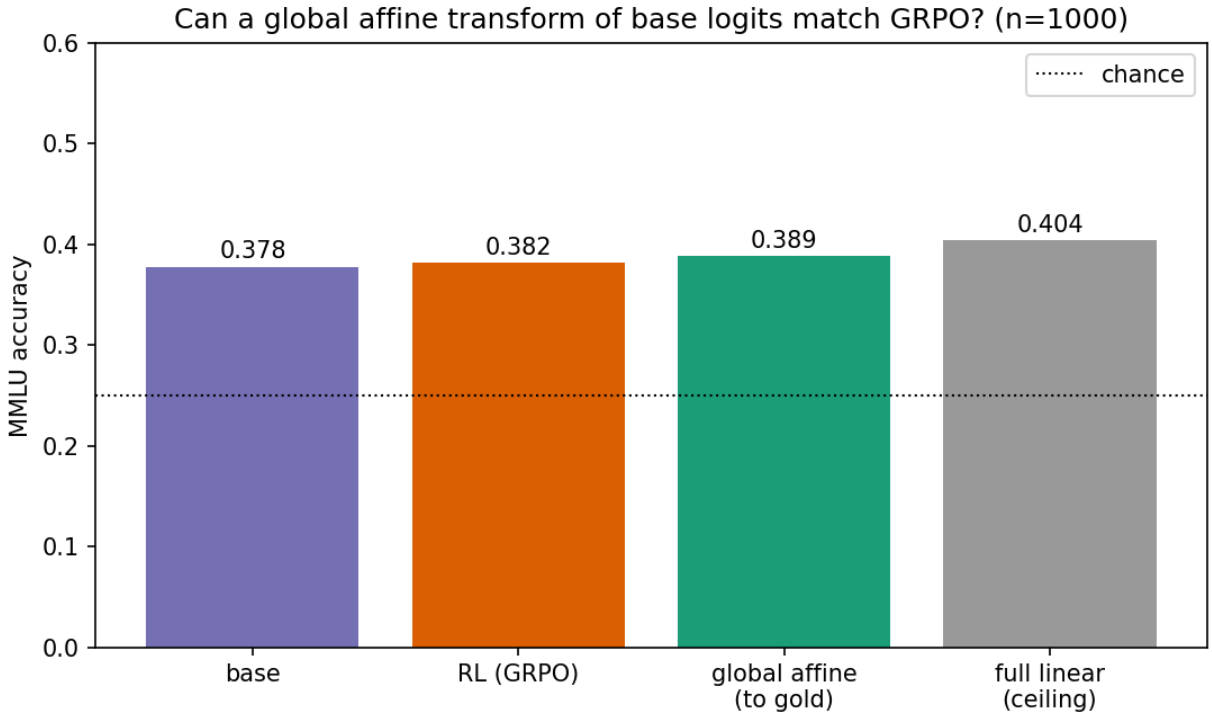

Figure 1: Figure 1. MMLU accuracy for the base model, the RL (GRPO) model, a global affine transform of base logits fit toward correctness, and a full linear map (ceiling). GRPO does not separate from base, and a global affine recalibration of base already matches or exceeds it.

GRPO produced no measurable accuracy gain. Base accuracy is 0.3780 (95 percent CI 0.3490 to 0.4090) and RL accuracy is 0.3820 (0.3520 to 0.4130), a difference of +0.0040 whose confidence intervals almost completely coincide, both well above the four-way chance level of 0.25. A paired test confirms the models are barely distinguishable: GRPO changed the predicted answer on only 27 of 1000 items (2.7 percent), split 12 base-wrong-to-RL-right against 8 base-right-to-RL-wrong, a McNemar chi-squared of 0.45 (p=0.50), so the direction of change is not significant. Because the denominator of the elicitation fraction is this near-zero gain, the fraction is ill-defined; the raw value (2.75) is an artifact of dividing by noise and we do not interpret it. For reference Figure 1 shows that a global affine transform of the base logits fit toward correctness reaches 0.3890 and a full linear map reaches 0.4040, but both are fit on the evaluation gold in cross-validation, so they are supervised recalibration ceilings and not a fair head-to-head with unsupervised GRPO; we report them only to bound what a static transform can do.

The mechanistic result is the robust one. A global affine transform explains 99.3 percent of the RL option-

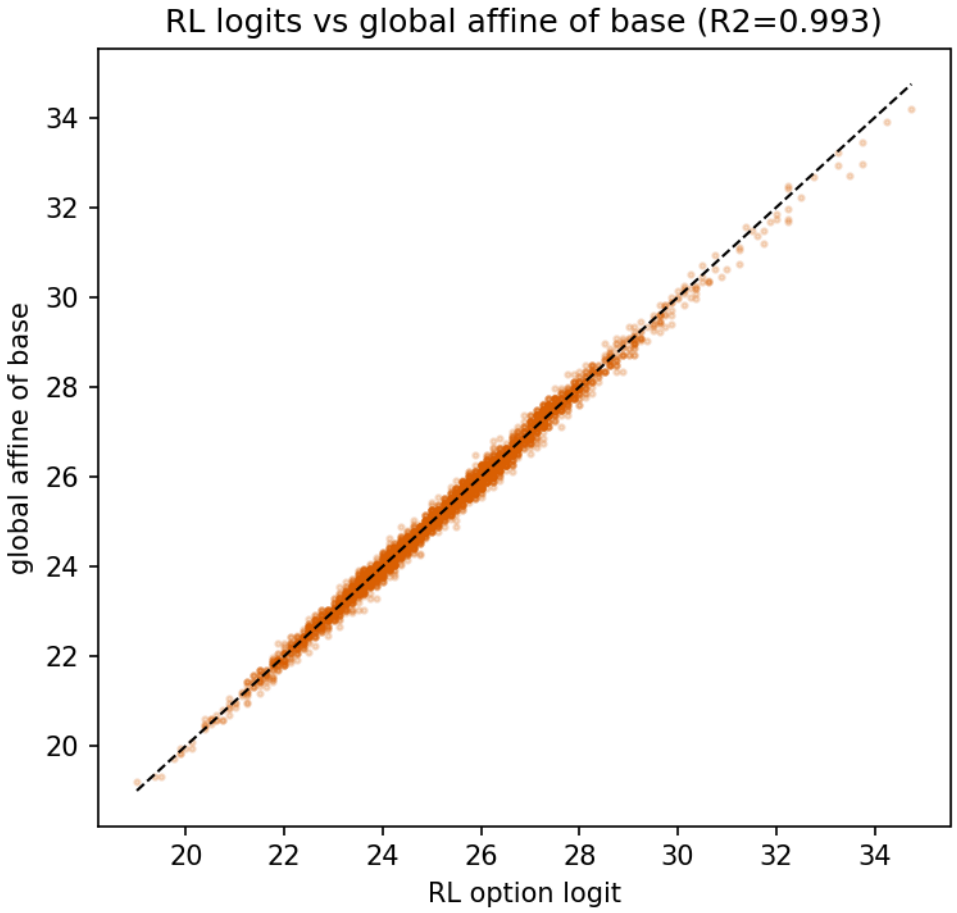

Figure 2: Figure 2. RL option logits against the global affine transform of base option logits. The fit lies near the identity line with R-squared 0.993.

logit variance (Figure 2), but the base logits already explain 98.5 percent with no transform at all, and a per-option bias with temperature fixed at one already reaches 99.3 percent. The fitted temperature is 0.989, essentially one, so the affine adds almost nothing beyond a constant per-option offset. In other words GRPO barely moved the logits (mean absolute change 0.180), and the small movement it made is, to within a fraction of a percent of variance, a per-option bias shift rather than any input-dependent restructuring. Figure 3 shows that shift is a near-constant upward push on every option, largest on option A (mean +0.271) and smallest on option D (+0.063), the signature of a mild label or position bias rather than improved discrimination.

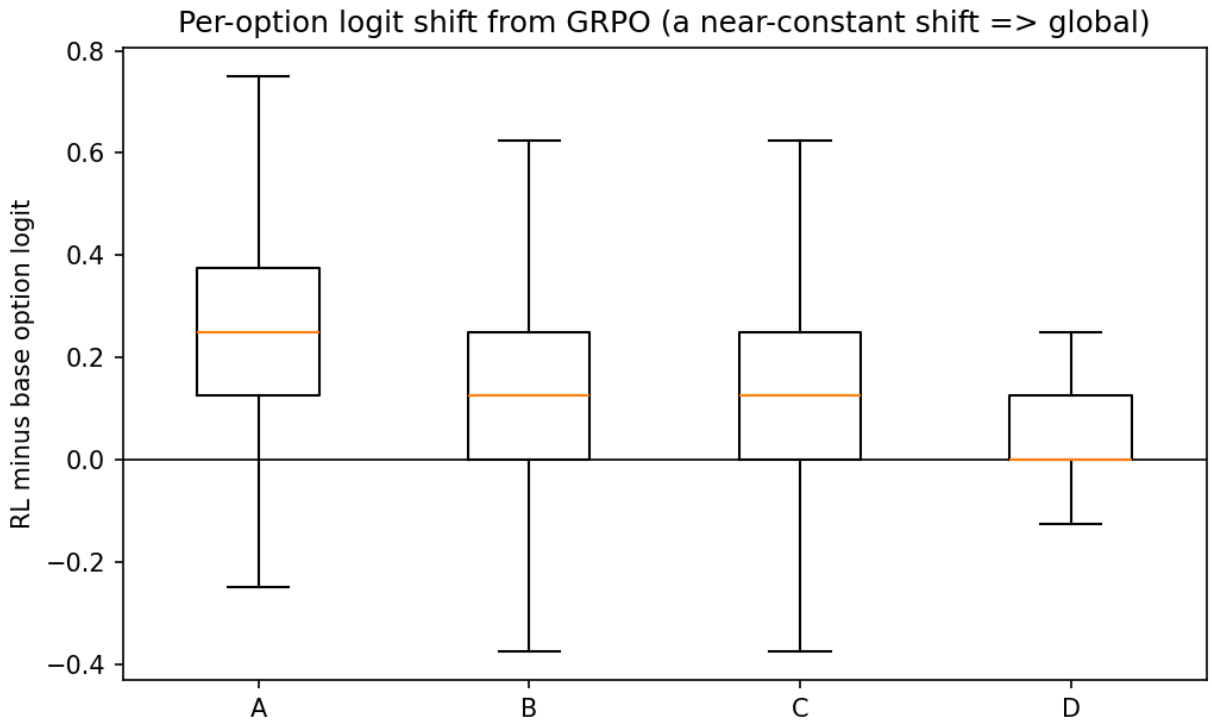

Figure 3: Figure 3. Per-option base-to-RL logit shift. Every option is pushed up by a near-constant amount, largest for option A, consistent with a global label-bias shift.

Figure 4 stacks the three variance-explained numbers and makes the reading direct: identity 0.985, bias-only 0.993, full affine 0.993. The model after GRPO is a global recalibration of the model before it, and a slight one.

4. Discussion and limitations

The honest reading is mostly about how little changed. GRPO moved the predicted answer on 2.7 percent of items with no significant direction and gained no accuracy, so the dominant fact is that the model barely changed. Conditional on that small change, its effect on the logits is almost entirely global: 99.3 percent of the base-to-RL logit variance is captured by a single temperature-and-bias transform, against 98.5 percent with no transform at all, so the affine adds only about one point of explained variance beyond the identity, and that increment is a per-option bias toward option A. We therefore do not claim a strong dichotomy of recalibration versus capability change; the data support the weaker and more defensible statement that the model barely moved and the little it moved is global. This is consistent with the elicitation view that post-training reweights rather than rebuilds [@zhou2023lima; @lin2023unlocking; @rafailov2023dpo], and it answers the literal question, a single global affine transform does recover GRPO's logit effect, but the headline framing of recovering an accuracy gain collapses because there is no gain to recover.

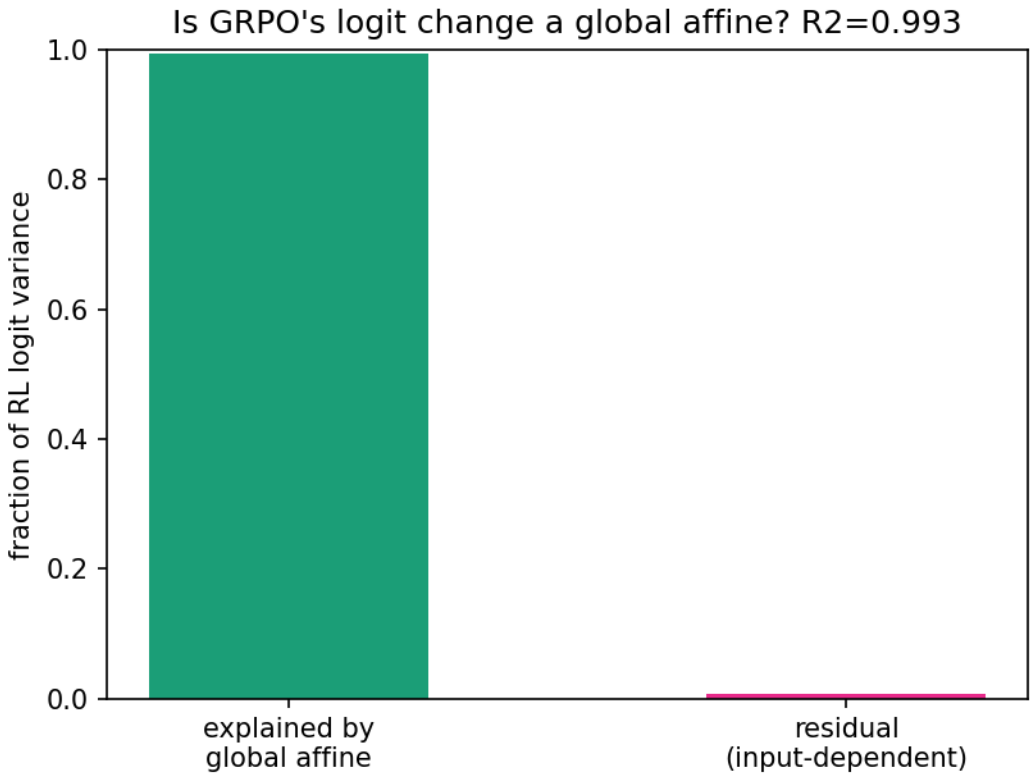

Figure 4: Figure 4. Variance of RL option logits explained by the identity (no transform), a per-option bias only, and the full global affine. The increments are small, so GRPO's change is both tiny and almost entirely global.

The limitations are decisive and we state them plainly. First, the GRPO treatment was weak: 120 steps on a 1.5 billion parameter model moved accuracy by +0.004, within noise, so the strong finding that the effect is global is partly the trivial fact that the model barely changed (identity already explains 98.5 percent of the variance); a treatment in the regime where outcome-only RL transforms reasoning [@guo2025deepseekr1] is the test that matters. Second, the scalar elicitation fraction we set out to measure is undefined here and we report it as such rather than as a number. Third, the supervised recalibration ceilings (affine-to-gold and full-linear) are fit on the evaluation labels in cross-validation, so they are not a fair head-to-head with unsupervised GRPO and we use them only to bound what a static transform can reach, not to claim a transform beats GRPO. Fourth, we report the R-squared and per-option-shift numbers as point estimates without confidence intervals, since bootstrapping them would require refitting the maps per resample; the paired prediction-level result (McNemar) is the uncertainty-aware backbone of the no-change claim, and the variance numbers are descriptive. Fifth, we study the four-option MMLU answer logits at a single forced-answer position, not generative reasoning, so the decomposition speaks to the answer distribution and not to the chain of thought. Sixth, this is one model family at one size with one reward. The constructive contribution is the affine-versus-identity-versus-full-linear decomposition as a cheap probe of whether an RL update is a global recalibration or something more, and the explicit specification that a credible elicitation-fraction measurement needs a treatment that produces a real accuracy gain.

5. Conclusion

We asked whether outcome-only GRPO's accuracy gain could be recovered by a single global affine transform of the base model's logits. On Qwen2.5-1.5B-Instruct there was no significant accuracy gain to recover (base 0.378, RL 0.382), so the scalar elicitation fraction is ill-defined. The mechanistic answer is nonetheless clean: GRPO changed the predicted answer on only 2.7 percent of items (McNemar p=0.50) and its effect on the logits is almost perfectly global, a per-option bias shift (largest on option A) that a temperature-and-bias transform reproduces to 99.3 percent of variance against a 98.5 percent no-transform baseline. At this scale outcome-only RL barely moved the answer distribution and what it moved was global, and the affine-versus-identity-versus-full-linear decomposition is a cheap diagnostic for telling a global recalibration from a real capability change at a strength where the question has teeth [@guo2025deepseekr1; @lightman2023letsverify].