Pluralistic Alignment via Self-Distillation from Synthetic User Feedback

Minsik Oh
Department of Computer Science
Stanford University
minsik@stanford.edu

Extended Abstract

Problem. Standard RLHF optimizes a single, population-averaged reward model and ships one "aligned" policy for every user. This collapses the diversity of human preferences (i.e. cultural / religious / philosophical / ethical priors) into a monolithic average that fits no individual well. We post-train an LLM steered towards a user's values and characteristics: a single policy that is steered to internalize an individual user's values via training-time textual feedback.

Method. We introduce **PAPO** (Pluralistic Alignment Policy Optimization), which replaces the scalar reward of RLHF with *natural-language critiques* from synthetic users. A training instance is a pair (x, u) of prompt x and persona u; the student $\pi_\theta(y \mid x, u)$ generates a response y, and a judge M conditioned on an itemized persona constitution p returns a critique $c \sim M(\cdot \mid x, u, p, y)$ explaining how y fails to match the persona. Following on-policy self-distillation (SDPO), we treat the feedback-conditioned policy $\pi_\theta(\cdot \mid x, u, c)$ as a self-teacher and distill its distribution back into the student, using a single *sequence-level* advantage $A(y) = \log \frac{\pi_\theta(y|x,u,c)}{\pi_\theta(y|x,u)}$ optimized with a clipped surrogate loss and a stop-gradient teacher. This trajectory-level credit was easier to stabilize than per-token logit-level distillation. We combine the self-distillation objective with GRPO via a weight β, and train Qwen3 8B, 4B, 1.7B models in the VeRL framework with GPT-5.5 as the persona judge.

Spec selection. We first run GRPO across a broad pool of alignment specs and rank them by difficulty (validation accuracy, `acc/mean@1`). Easy specs saturate near 100% and leave no headroom to show a training effect, so we select the low-performing specs where richer feedback should help most. This is especially important as a threshold is utilized to determine whether feedback should be applied.

Results. PAPO improves over GRPO on every selected spec.

Persona Spec	Baseline (8B)	GRPO	PAPO (ours)	Δ
Zen Buddhist Koans & Contemplation	20%	70%	**80%**	+10%
Rationalism & Bayesian Reasoning	4%	44%	**82%**	+38%
Progressive Policy Advocacy	0%	0%	**56%**	+56%

Table 1: Accuracy of PAPO over unaligned baseline and GRPO. Full specs defined in Section 3.1.

Analysis. Across training, three metrics move together: more on-policy samples clear the success threshold, the feedback-used fraction falls, and critic score rises. The model is learning to satisfy the persona *without* a critique; the preference internalization behavior the objective targets. We find that early feedback consumption sets up the later gains.

Conclusion. We present PAPO, a method that turns natural-language critiques from synthetic user personas into a self-distillation signal for personalization. On specs where standard GRPO struggles, PAPO recovers substantial alignment from 10% to 56% and its training dynamics suggest the policy internalizes persona preferences rather than depending on critiques at inference. This points toward a scalable route to pluralistic alignment that needs no per-user human response data.

Preprint.

Abstract

Standard RLHF optimizes a single, population-averaged reward and ships one aligned policy for every user, collapsing the diversity of human preferences into an average that fits no individual well. This gives rise to pluralistic alignment, which prioritizes individual user's values and characteristics by incorporating their population-specific leanings to alignment goals. We introduce **PAPO (Pluralistic Alignment Policy Optimization)**, which integrates natural-language critiques from synthetic user personas. We first stabilize SDPO (Self-Distilled Policy Optimization) training via selecting sequence-level advantage formulation and performing GRPO simultaneously. Then, a persona-conditioned judge critiques each response, and the feedback-conditioned policy acts as a self-teacher whose distribution is distilled into the student. Selecting persona specs where GRPO struggles (Zen Buddhist, Bayesian Reasoning, Progressive Policy), PAPO improves training stability and validation accuracy on every spec from 10% to 56%. As training proceeds, successfully judged samples rise while feedback usage falls, indicating that persona preferences are internalized into the policy. Thus, we validate that synthetic personalized feedback can lead to effective steering towards successful pluralistic alignment. PAPO offers a scalable path to pluralistic alignment that does not require human response data.

1 Introduction

Large language models are increasingly deployed as general-purpose assistants serving millions of users with vastly different backgrounds, values, and expectations. The dominant recipe for shaping their behavior is reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022): human annotators express pairwise preferences over model outputs, a reward model is fit to those preferences, and the policy is optimized against that reward. This pipeline has been remarkably effective at making models more helpful and harmless on average, and is now standard across frontier systems.

Yet "on average" is precisely the limitation. A single reward model trained on aggregated human preferences encodes a population-level consensus, and the resulting policy is tuned to satisfy an abstract median user. Recent work argues that this aggregation does not merely fail to capture diversity but actively *reduces* pluralism, flattening the range of legitimate human values a model might reasonably reflect (Sorensen et al., 2024). The values that matter for personalization (how formal or terse to be, how much detail to give, which cultural / religious / philosophical / ethical priors to foreground, how to interpret an ambiguous request) are exactly the dimensions a population average erases.

Our objective is to develop a reinforcement learning method that post-trains a large language model (LLM) toward *pluralistic alignment via personalization*: at inference time, the same model should be steerable to align in-depth with the preferences of an individual user. The key idea is to combine self-distillation (Hübotter et al., 2026; Zhao et al., 2026) with *natural-language critiques* produced by *synthetic users*; LLMs prompted to role-play diverse user personas with distinct preferences. Natural-language critiques are very rich source of information: they localize what a user disliked and suggest how to fix it, providing rich supervision well-suited to learning personalized outputs.

Our contributions are following:

1. We are, to our knowledge, the first to apply self-distillation from text feedback to *pluralistic alignment* rather than domains with verifiable correctness.

2. We replace human/expert judges with *persona-conditioned LLM judges* and study whether self-distillation can result in an optimized personalized policy.

3. We report whether the textual feedback has been applied during training time via analyzing *feedback usage rates* and samples.

2 Related Work

RL from text feedback. Song et al. (2026) formalize Reinforcement Learning from Text Feedback (RLTF) and propose two methods: Self Distillation (RLTF-SD), which distills feedback-conditioned second-turn rollouts into the single-turn policy, and Feedback Modeling (RLTF-FM), which adds an auxiliary loss predicting critiques. They show consistent gains over GRPO (Shao et al., 2024) on reasoning, math, and creative writing. However, their setting targets verifiable correctness and treats feedback as a generic improvement signal; not as a vehicle for encoding *user identity*, and their judges are task-grading experts rather than preference-bearing personas.

Self-distillation policy optimization. Hübotter et al. (2026) introduce SDPO, which uses the current policy conditioned on rich environment feedback as a self-teacher and distills its feedback-informed next-token distribution back into the student via a logit-level KL loss. SDPO yields dense credit assignment and outperforms GRPO on code generation, scientific reasoning, and tool use. Yet SDPO is studied exclusively in verifiable-reward domains (LeetCode-style runtime errors, unit tests); it has not been applied to subjective, user-centric preferences where "correctness" is intrinsically pluralistic.

Pluralistic alignment and personalization. Sorensen et al. (2024) lay out a roadmap distinguishing Overton, steerable, and distributional pluralism, and argue that current alignment procedures actively reduce pluralism. Chakraborty et al. (2024) prove an impossibility result for single-reward RLHF and propose a mixture-of-reward-models with a MaxMin objective. Jang et al. (2023) introduce RLPHF, decomposing preferences into multi-objective dimensions and merging per-dimension policies post-hoc. Hindsight-style methods condition policies on feedback as a goal (Liu et al., 2023; Zhang et al., 2023). These approaches either collapse diversity into scalars/dimensions (losing the very pluralism they aim to capture), require multiple policies to be merged, or condition on personas only at inference without an explicit training-time mechanism that teaches the model *how* to internalize preferences. None combine *synthetic-user textual critiques* with *self-distillation-based credit assignment* to produce a single personalizable policy.

Gap. Existing methods either (a) leverage rich feedback in verifiable, single-correct-answer domains that do not transfer to personalization, yielding generic alignment (Song et al., 2026; Hübotter et al., 2026), or (b) target pluralism but lack dense, feedback-derived credit assignment, relying on expensive human preferences and/or yielding weak alignment (Sorensen et al., 2024; Chakraborty et al., 2024; Jang et al., 2023). We address both gaps simultaneously.

3 Method

3.1 Synthetic Environment

Our training data and evaluation constitutions were synthetically generated using large (120B+) open-source LLMs. 9 persona specs were selected via querying an LLM according to clarity of value systems and estimated difficulty. For each selected specs, approx. 200 constitutions generated and only constitutions with "critical" priority was retained for persona judge. Then, 10,000 questions were synthetically generated for each selected spec, from which 500 questions were selected for training and evaluation (9:1 ratio).

A mix of "Qwen/Qwen3.5-122B-A10B" and "openai/gpt-oss-120b" were used for constitution generation, from each 100 constitutions were generated per a given persona spec. "Qwen/Qwen3.5-397B-A17B-FP8" was used for merging 2constitutions, deduplication and question generation. This synthetic data was internally prepared before the course (last quarter) for a pending publication, thus unrelated to the course. "Qwen/Qwen3" models of 3 sizes (8B, 4B, 1.7B) were trained with GRPO and PAPO using VeRL framework in an effort to replicate SDPO setting.

3.2 Persona Spec Selection

We first run GRPO across 9 candidate alignment specs and rank them by difficulty: the converged validation accuracy (`acc/mean@1`). Easy specs saturate near 100% and leave no headroom to demonstrate a training effect. We therefore select 3 **low-performing** specs, where GRPO struggles and self-distillation has room to improve. Table 4 shows the result of GRPO training.

Following specs were chosen:

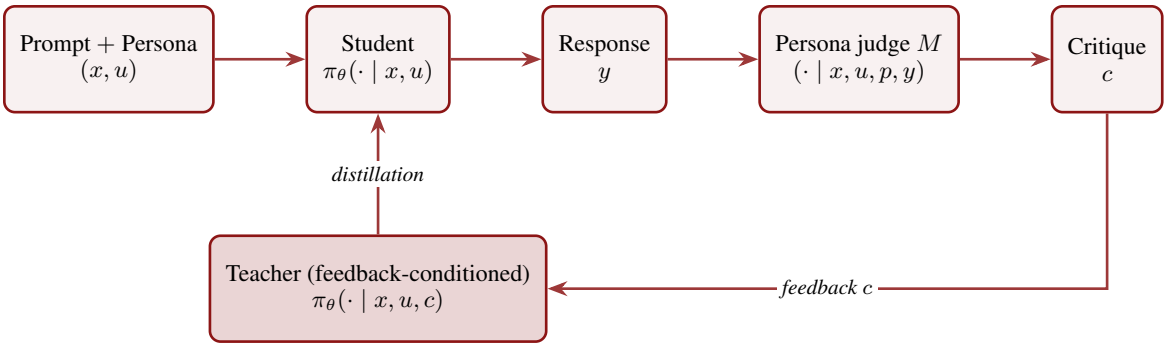

Figure 1: Synthetic-user loop. The persona judge turns a response into a natural-language critique; the feedback-conditioned policy acts as a self-teacher that is distilled back into the student. Description at Section 3.3.

1. a model for zen buddhist koans and contemplation (shortened as Zen Buddhist)
2. a model grounded in rationalism and bayesian reasoning (shortened as Bayesian Reasoning)
3. a model for progressive policy advocacy (shortened as Progressive Policy)

Experimental results of this selection process is reported in Section 5.0.1.

3.3 Self-Distillation of Persona Critique

A training instance is a tuple (x, u) where x is a user prompt and u is a persona description [1]. The policy $\pi_\theta(y \mid x, u)$ generates a response y. A model M conditioned on itemized personal constitutions p acts as an LLM-judge, producing a textual critique $c \sim M(\cdot \mid x, u, p, y)$ that explains, in natural language, how y fails to match the persona's preferences.

We differ from SDPO in that synthetic persona judge M generates textual critique rather than using feedback from verifiable environments such as runtime error messages. Whether persona judge M can provide reliable feedback is critical to our algorithm, for which we see promising results.

3.4 Training Stability Improvements

We treat the feedback-conditioned policy $\pi_\theta(\cdot \mid x, u, c)$ as a self-teacher and distill its distribution into the student $\pi_\theta(\cdot \mid x, u)$. Rather than a noisy per-token signal, we assign a **single sequence-level** advantage to the whole rollout y:

$$A(y) \;=\; \log \frac{\pi_\theta(y \mid x, u, c)}{\pi_\theta(y \mid x, u)} \;=\; \sum_t \log \frac{\pi_\theta(y_t \mid x, u, c, y_{<t})}{\pi_\theta(y_t \mid x, u, y_{<t})},$$

optimized with a clipped surrogate loss, with the teacher as a stop-gradient target. This trajectory-level credit was easier to stabilize than per-token logit-level distillation in SDPO. Finally, we **combine GRPO and self-distillation** loss using β weight for latter.

4 Experimental Setup

We experimented on Qwen3 8B, 4B and 1.7B models with VeRL framework. Judge model used is GPT-5.5. We implemented persona data processing scripts, persona judge M and GRPO & SDPO dual loss. We ablate over 4 hyperparameters: KL Divergence / Jensen-Shannon Divergence interpolation α, Self-distillation loss ratio β, judge success threshold, and teacher update rate. We found that PAPO is robust to α values, thus we set it to 0.5. We found that PAPO improves upon GRPO when β is sufficiently high, while setting it too high didn't make much difference. Thus we set it to 0.5. We ablate upon persona-judge success threshold and teacher update rate and report the outcome in Section 5.0.2. In contrast to SDPO where default persona-judge success rate was 0.7,

[1] e.g., a Scandinavian, a Stoic philosophy expert, a zen Buddhist monk

Hyperparameter	Candidates	Chosen
KLD/JSD interpolation α	$0,\ 0.5,\ 1.0$	0.5 (JSD)
Self-distillation loss ratio β	$0.1,\ 0.5,\ 1.0$	0.5
Judge success threshold	$0.7,\ 0.9,\ 0.95$	$0.9,\ 0.95$
Teacher update rate	$0.1,\ 0.25,\ 0.5$	$0.1,\ 0.25$

Table 2: Hyperparameter selection process. Description in Section 4. Chosen hyperparameters are ablated in Section 5.0.2.

Persona Spec	Baseline	GRPO	PAPO (ours)	Δ
8B				
Zen Buddhist Koans & Contemplation	20%	70%	**80%**	+10%
Rationalism & Bayesian Reasoning	4%	44%	**82%**	+38%
Progressive Policy Advocacy	0%	0%	**56%**	+56%
4B				
Zen Buddhist Koans & Contemplation	20%	86%	**96%**	+10%
Rationalism & Bayesian Reasoning	2%	42%	**58%**	+16%
Progressive Policy Advocacy	0%	0%	**44%**	+44%
1.7B				
Zen Buddhist Koans & Contemplation	14%	58%	**86%**	+28%
Rationalism & Bayesian Reasoning	2%	**14%**	**14%**	+0%
Progressive Policy Advocacy	0%	0%	**4%**	+4%

Table 3: Final, optimal validation accuracy of PAPO over unaligned baseline and GRPO across model scales (8B, 4B, 1.7B). Full specs defined in Section 3.1. Discussion in Section 5.

we found that larger values work better. Similarly, default teacher update rate was 0.05 for SDPO but larger values worked; however too large value of 0.5 degraded performance. See Table 2 for hyperparameter choices. We ablate upon chosen hyperparameters in Section 5.0.2.

5 Quantitative Results

We report our headline result in Table 3. We find that our PAPO method can improve upon GRPO models of varying difficulty, ranging from 0% to 70%. Performance increases of 10% to 56% is observed. GRPO itself tends to be effective to a degree as well, but PAPO can recover failed unstable training runs such as for progressive policy spec. Performance increases is higher if initial performance was lower. We observe performance increases across all model sizes, with small models sometimes aligning better even when initial performance is low. This may be due to small models being easier to overfit to a specific criteria.

5.0.1 Persona Spec Selection

Before applying PAPO, we use GRPO to establish a baseline and to identify which persona specs can be improved (methodology in Section 3.2). We train a separate GRPO policy for each of nine candidate specs and track validation accuracy (`val-core/persona/acc/mean@1`) and training stability (`critic/score/mean`); Table 4 reports the pre- and post-training accuracy and Figure 2 shows the per-spec training process.

We find that GRPO is broadly effective, but the difficulty of a spec varies enormously. Most specs improve substantially under GRPO, and the *starting* accuracy of the untrained 8B model is the clearest predictor of how a spec behaves. Specs the base model already handles well (ommunitarianism and natural law theory) are nearly saturated at initialization and converge to 95%+ with little headroom to learn. A second group begins low to moderate (social democratic governance, globalist multilateral policy, Jain philosophy, centrist pragmatic policy) but is readily learnable, with GRPO lifting all

Persona Spec	Baseline	GRPO
Communitarianism	80%	100%
Natural Law Theory	86%	96%
Globalist Multilateral Policy	42%	94%
Social Democratic Governance	34%	92%
Jain Philosophy	22%	86%
Centrist Pragmatic Policy	26%	78%
Zen Buddhist Koans & Contemplation	**20%**	**70%**
Rationalism & Bayesian Reasoning	**4%**	**44%**
Progressive Policy Advocacy	**0%**	**0%**

Table 4: Per-persona validation accuracy (`val-core/persona/acc/mean@1`) before (Baseline 8B model) and after GRPO training. Specs with low GRPO accuracy (< 75%) were selected for our PAPO experiments (bolded). This ensures PAPO can improve the model even with a success threshold on responses that control feedback applicability. Discussion in Section 5.0.1. Training trends in Fig. 2.

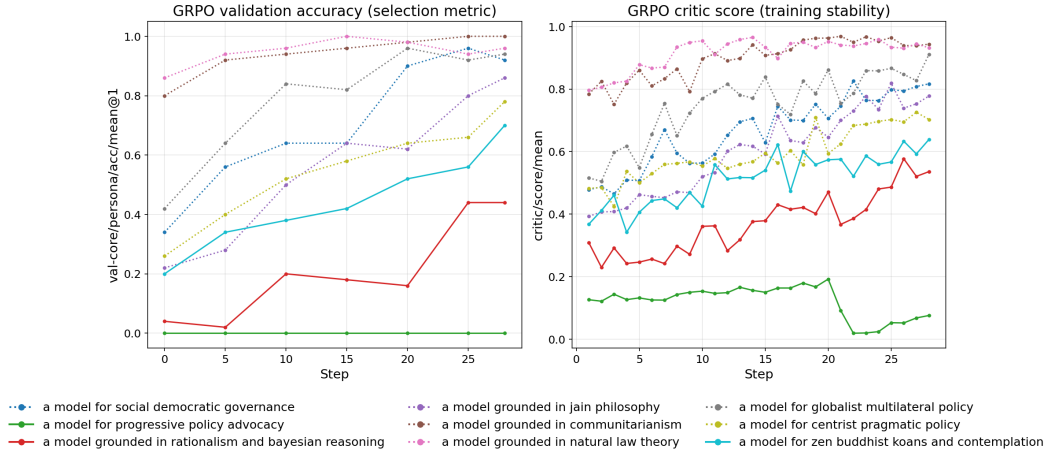

Figure 2: Persona spec selection GRPO experiments. We observe clear trends in both validation accuracy and critic score. Bottom 3 persona specs were selected (others in dotted lines) as they have most room for improvement with PAPO. Discussion in Section 5.0.1. Final numbers in Table 4.

of them to 78-94%. For these specs the limited room between baseline and ceiling makes them poor testbeds for a new method. For high scoring specs, this is further aggravated by the fact that a persona-judge success threshold of 90%+ determines whether a feedback will be utilized for the response. If feedback is not applied due to the threshold, it cannot improve the model.

Most specs show a rising critic score that tracks its accuracy gains, proving that training stability is maintained during model training. However, progressive policy model stays flat near 0.15 for most of training and then *collapses* after roughly step 20, falling toward zero rather than improving. This means training for progressive policy model was unstable. Upon sample analysis, it was determined that the failing model tends to generate long, truncated paragraphs that tries to cover both "progressive" and "conservative" viewpoints rather than aligning to intended progressive policy viewpoint. Integrating natural language feedback will stabilize the training as we will soon report.

5.0.2 Hyperparameter Ablations

As discussed in Section 4, we fix α and β values and toggle persona-judge success threshold and teacher update rate for our experiments. We sweep both on the Zen Buddhist spec, judge success threshold $\in \{0.9, 0.95\}$ and teacher update rate $\in \{0.10, 0.25\}$. Figure 3 reports validation accuracy and critic score for all four configurations against the GRPO baseline. We observe 2% to

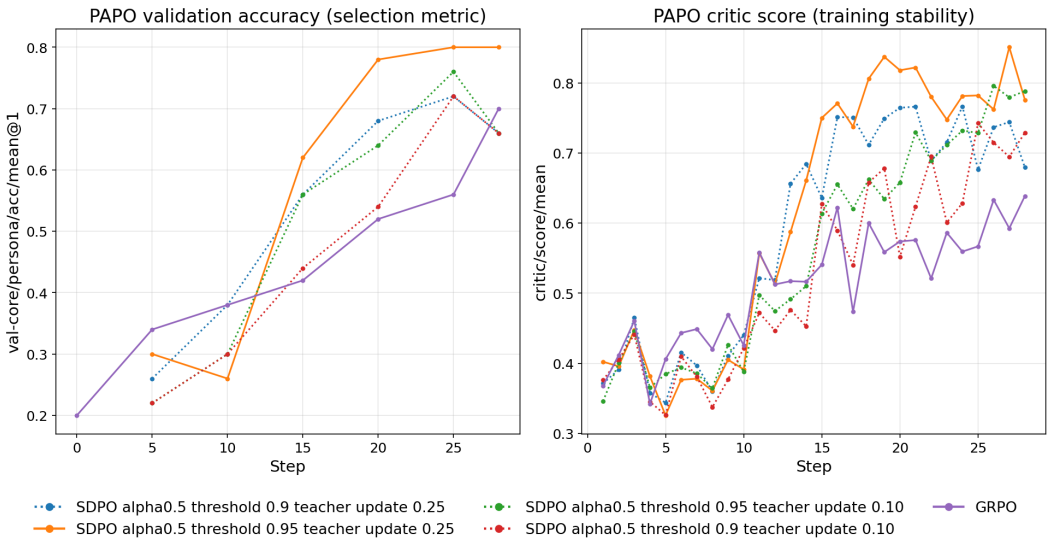

Figure 3: Hyperparameter ablation on the Zen Buddhist Koans & Contemplation spec, sweeping success threshold $\in \{0.9, 0.95\}$ and teacher update rate $\in \{0.10, 0.25\}$, against the GRPO baseline. We observe consistent improvements of 2% to 10% across configurations, with best, clearly improved model having success threshold of 0.95 and teacher update rate of 0.25 (other configurations dotted, except GRPO baseline). This validates the statistical and configurable robustness of our algorithm. Discussion in Section 5.0.2.

10% improvements across models, proving the robustness of the algorithm to hyperparameters and repeated runs.

The combination of the highest persona-judge success threshold 0.95 and the larger teacher update rate 0.25 is the clear winner: it pulls away from the other runs and converges to 80% validation accuracy, the best of any configuration. This trend extends to training stabililty where it's consistently highest. Notably, the other 0.95-threshold run (with the slower teacher update rate 0.10) also achieves 5% performance increase, in comparison to 0.9 threshold runs that achieve only 2% performance increase. This might mean that 0.95 threshold enables more stable training by incorporating the judge feedback more.

6 Qualitative Analysis

6.1 Is feedback used? Why self-distillation works.

As training progresses, more on-policy responses clear the success threshold, requiring no feedback. Thus the model is learning to satisfy the persona without the critique, which reflects the internalization of preferences to the policy. We illustrate this in Figure 4. We find that early feedback consumption sets up the later gains.

The three panels move in a consistent and interpretable way. As training proceeds, more on-policy samples clear the threshold (success fraction rises) and the feedback-used fraction correspondingly falls: once the student can satisfy the persona on its own, the teacher's critique-conditioned rollout no longer improves on it, so feedback is invoked less often. Critic score rises in tandem, meaning that training is stable. The model is learning to satisfy the persona *without* a critique, internalizing the preference into the policy rather than depending on feedback at inference.

7 Discussion

Natural-language critiques from synthetic personas can be turned into a usable training signal for personalization via self-distillation, and on specs where standard GRPO struggles, PAPO delivers clear gains even when GRPO is unstable (i.e. progressive policy advocacy). The benefit is largest

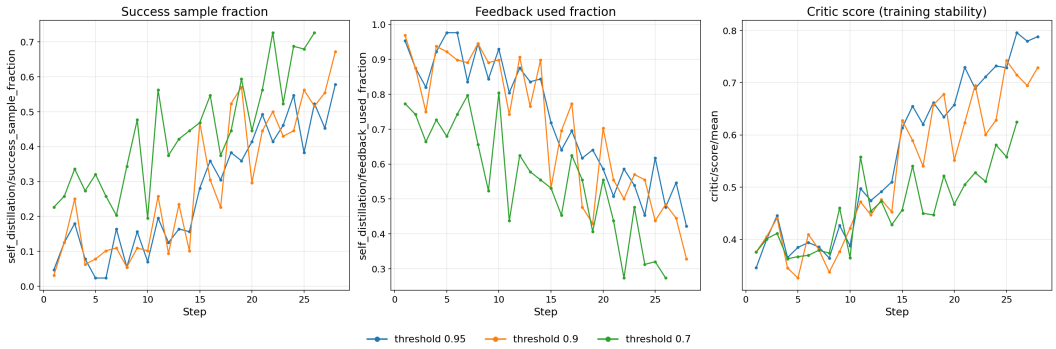

Figure 4: Training dynamics of persona-judge success threshold ablation on the Zen Buddhist Koans & Contemplation spec, sweeping success threshold $\in \{0.7, 0.9, 0.95\}$ while fixing teacher update rate to 0.10. Within each run, sample success fraction rises while feedback usage falls. This means the personalized preference of the dataset is internalized, requiring less feedback as more samples pass the persona-judge success threshold. This is consistent among three runs, validating statistical robustness. Discussion in Section 6.1.

where the baseline is weakest, but still achieves respectable 10% performance increase on initial 70% specification (i.e. Zen Buddhist). We leave it to future work on whether PAPO can be applied to easy specs that already reach 90% only with GRPO.

The feedback usage analysis clarifies why. The declining feedback-used fraction over training is genuine evidence of internalization, in that the student increasingly satisfies the persona-judge without needing a critique but only when the acceptance bar is high. A permissive threshold produces the same falling feedback curve for the opposite reason: it declares mediocre rollouts successful, stops soliciting corrective feedback early, and plateaus at lower quality. The lesson is that "less feedback used" is a desirable signal only when "success" means genuine persona alignment, which is what a strict threshold enforces.

8 Conclusion

We introduced PAPO, a method that converts natural-language critiques from synthetic user personas into a sequence-level self-distillation signal for pluralistic alignment. By selecting persona specs that resist standard GRPO and training against persona-conditioned critiques, PAPO improves alignment on difficult-to-align specs and can improve training stability. Its training dynamics indicate that the policy internalizes persona preferences, satisfying them increasingly without a critique present, which is the behavior a personalized model must exhibit at inference. Our analysis shows this signal is trustworthy under a strict success threshold. Together these results offer a scalable route toward pluralistic alignment that requires no per-user human response data. We leave it to future work on whether human-in-the-loop methods can be developed that improve further on the methodology.

9 Team Contributions

- **Minsik Oh**: Planned the project, prepared the data, implemented the algorithms, executed experiments and analyzed results.

References

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Alignment with diverse human preferences, 2024. URL `https://arxiv.org/abs/2402.08925`.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation, 2026. URL `https://arxiv.org/abs/2601.20802`.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging, 2023. URL https://arxiv.org/abs/2310.11564.

Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. Chain of hindsight aligns language models with feedback, 2023. URL https://arxiv.org/abs/2302.02676.

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.

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.

Yuda Song, Lili Chen, Fahim Tajwar, Remi Munos, Deepak Pathak, J. Andrew Bagnell, Aarti Singh, and Andrea Zanette. Expanding the capabilities of reinforcement learning via text feedback, 2026. URL https://arxiv.org/abs/2602.02482.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. A roadmap to pluralistic alignment, 2024. URL https://arxiv.org/abs/2402.05070.

Tianjun Zhang, Fangchen Liu, Justin Wong, Pieter Abbeel, and Joseph E. Gonzalez. The wisdom of hindsight makes language models better instruction followers, 2023. URL https://arxiv.org/abs/2302.05206.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.