# OVERTON PLURALISTIC REINFORCEMENT LEARNING FOR LARGE LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Existing alignment paradigms remain limited in capturing the pluralistic nature of human values. Overton Pluralism (OP) addresses this gap by generating responses with diverse perspectives from a single query. This paper introduces **OP-GRPO** (Overton Pluralistic Group Relative Policy Optimization), a reinforcement learning framework for *implicit* Overton Pluralism that enables a single LLM to produce pluralistic responses without explicit prompting or modular orchestration. Our workflow consists of two main steps: 1) *Similarity estimator training*, which fine-tunes a Sentence Transformer for OP tasks to provide a more accurate coverage evaluation of the given responses; and 2) *OP-GRPO training*, which incorporates this similarity estimator into a carefully designed dual-reward system to ensure both broad coverage of genuine human perspectives and the uniqueness of each perspective, thereby promoting diversity. Empirical results demonstrate that OP-GRPO achieves a *"Small models, Big perspective coverage"* effect: our trained `Qwen2.5-3B-Instruct` surpasses the `GPT-OSS` (20B) baseline with a 37.4% relative accuracy gain on the Natural Language Inference (NLI) benchmark. It also outperforms a modular-architecture baseline with a 19.1% relative improvement. Evaluations with `GPT-4.1` as LLM judge for response quality assessment further confirm the robustness of our approach.

## 1 INTRODUCTION

Recent advances in AI alignment have significantly improved the quality of responses generated by large language models (LLMs) (Yu et al., 2025; Ji et al., 2025). Techniques such as supervised fine-tuning (Harada et al., 2025; Fan et al., 2025), pre-training (Tack et al., 2025), and reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) have been introduced to ensure more close alignment of these models with human values, intentions, and preferences (Leike et al., 2018). However, existing methods commonly optimize towards a singular consensus objective. This oversimplification would inadvertently marginalize minority perspectives, amplify dominant viewpoints disproportionately, and generate outputs that are contentious, biased, or brittle (Khalifa et al., 2020; Casper et al., 2023; Kirk et al., 2023). Such challenges arise due to the inherent pluralism of human values, which vary substantially across communities, cultures, demographics, and individual perspectives (Sorensen et al., 2024a). This limitation highlights a critical need to explore novel methods aimed at preserving and improving output diversity during training.

To address this issue, the concept of *Overton pluralism (OP)* has been proposed (Sorensen et al., 2024c). Overton pluralism advocates explicitly surfacing multiple defensible perspectives within the *Overton window* (Oxford English Dictionary, 2023), the spectrum of socially acceptable or viable viewpoints at a specific time, rather than compressing diverse stances into a single canonical response (Sorensen et al., 2024c). Achieving Overton pluralism has numerous benefits: it enhances transparency by explicitly articulating reasons behind reasonable disagreements, reduces hidden biases by ensuring diverse views are represented, and facilitates downstream decision-making by empowering users and policymakers to select perspectives closely aligned with their values.

This paper introduces a novel *implicit Overton pluralism* framework that, for the first time, leverages Reinforcement Learning from Human Feedback (RLHF), specifically employing the Group Relative Policy Optimization (GRPO) algorithm (Shao et al., 2024), to optimize training policies that guide LLMs toward **Overton pluralism**, as shown in Figure 1. The methodology consists of two

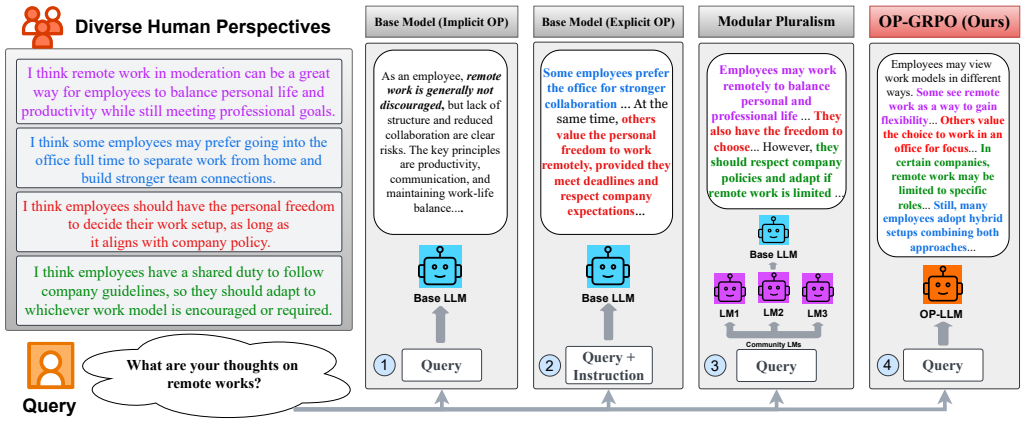


Figure 1: **An overview of different examples aimed at generating Overton Pluralism (OP) windows.** On the top left corner are real multi-perspective responses from diverse human groups to a given query. The figure then illustrates: (1) responses generated by a base LLM using an implicit OP prompt; (2) the same base LLM prompted with an explicit OP instruction; (3) a modular pluralism approach that combines outputs from different community-specific LLMs, summarized by a final LLM to form the OP window (Feng et al., 2024); and (4) our RL-trained pipeline, which achieves the highest correct coverage of human reference perspectives in OP responses.

main components: first, fine-tuning a Sentence Transformer (SBERT) model (Reimers & Gurevych, 2019) with hyperparameter optimization to adapt it to the Overton pluralism task; second, using the tuned SBERT model with the Mutual-Best Greedy Matching (MBGM) strategy as the foundation for a dual pluralistic reward system to train Overton pluralism policies with the GRPO algorithm. This approach enables even relatively small models to naturally generate Overton windows containing diverse viewpoints that more accurately reflect human perspectives. Empirical evaluation demonstrates the effectiveness of this approach. Even compact 1.5B and 3B models under OP–GRPO achieve state-of-the-art performance on NLI and LLM-as-Judge benchmarks (Feng et al., 2024; Lake et al., 2024), while also surpassing Modular Pluralism, an AI system can generate Overton pluralistic responses (Feng et al., 2024). Crucially, the OP–GRPO framework is designed to mitigate the loss of perspectives observed in conventional alignment methods (Sorensen et al., 2024c). These results confirm that the approach not only advances alignment with pluralistic values but also establishes a scalable, efficient, and practical pathway toward Overton Pluralism–aware language systems.

## 2 Related Work

Research on pluralistic alignment in LLMs spans training-phase methods that embed heterogeneous preferences into model parameters (Chen et al., 2024a; Xu et al., 2024; Harland et al., 2024; Srewa et al., 2025), and inference-time strategies that dynamically steer outputs (Chen et al., 2024b; Adams et al., 2025; Klassen et al., 2024; Caputo, 2024; Vamplew et al., 2024). In parallel, Sorensen et al. (2024c) first formalized the concept of *Overton Pluralism* by emphasizing the need for models to surface multiple defensible perspectives within the socially acceptable window. Follow-up work such as *ValuePrism* and *Kaleido* offered high-quality datasets and evaluation tools to encode and assess human values across contexts (Sorensen et al., 2024b). Lake et al. (2024) highlighted that alignment should move from distributional to Overton Pluralism. These contributions established Overton pluralism as both a conceptual paradigm and a practical challenge for scalable alignment.

Closest to our work, Modular Pluralism (Feng et al., 2024) operationalized Overton pluralism by orchestrating a base LLM with community-specific models to synthesize diverse perspectives. However, it also relies on an additional summary model to merge these outputs into a coherent Overton window. While effective, this approach requires multi-model collaboration at inference, introducing complexity and latency. In contrast, our work introduces an implicit OP framework trained with RLHF, enabling even a single small model to learn pluralistic behavior directly.

# 3 PRELIMINARIES.

**Reinforcement Learning from Human Feedback (RLHF).** RLHF casts alignment as learning a policy $\pi$ that maximizes rewards derived from human preferences. For a prompt $x$ and response $a$, the reward $r(x, a)$ guides the objective

$$\mathcal{J}(\pi) = \mathbb{E}_{x\sim\mathcal{X},\, a\sim\pi}\left[r(x,a) - \tau D_{\mathrm{KL}}\big(\pi(\cdot|x) \,\|\, \pi_{\mathrm{ref}}(\cdot|x)\big)\right], \tag{1}$$

where $\pi_{\mathrm{ref}}$ is a reference policy and $\tau > 0$ regulates deviation. The KL term ensures $\pi$ stays close to $\pi_{\mathrm{ref}}$, preserving useful behavior while preventing drift. Since direct human rewards are impractical to encode, learned reward models approximate $r(x, a)$ using preference data (Ouyang et al., 2022).

**Sentence Transformer (SBERT).** Sentence Transformers (Reimers & Gurevych, 2019) extend BERT encoders (Devlin et al., 2019) to produce fixed-length embeddings for semantic similarity tasks. By mapping text into a continuous vector space, SBERT enables efficient similarity measurement via cosine similarity $s(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$, which captures relational meaning independent of vector magnitude. This makes it valuable for preprocessing, reward shaping, and evaluation in OP tasks. Unlike mathematical reasoning, where correctness can be verified directly (Zuo et al., 2025; Dai et al., 2025; Dong et al., 2025), OP fine-tuning demands nuanced reward assignment for pluralistic responses. SBERT addresses this by providing semantically meaningful embeddings that support reliable similarity-based reward modeling.

**Overton Pluralistic Window.** The *Overton window* denotes the set $W(t)$ of perspectives considered socially acceptable at time $t$ (Oxford English Dictionary, 2023). A perspective $p$ is admissible iff $p \in W(t)$, with $W(t)$ shifting over time as norms and values evolve. *Overton pluralism* extends this idea by requiring models to present not a single canonical answer, but a set of defensible perspectives $Y = \{y_1, \ldots, y_k\} \subseteq W(t)$ (Sorensen et al., 2024c). These perspectives should be structured to indicate applicability, trade-offs, and uncertainties, and should reflect viewpoints from different social groups. Thus, Overton pluralism embeds the Overton window into alignment, ensuring AI outputs remain both diverse and socially grounded.

# 4 METHODOLOGY

## 4.1 DATASET PREPARATION

**OP Dataset Refinement.** To support fine-tuning for Overton pluralism, we begin with the ValuePrism dataset (Sorensen et al., 2024b), which contains 218k social situations paired with machine-generated perspectives, later re-examined by humans. We reformat these into multi-perspective templated sentences to construct the **OP-V1** dataset. In order to address redundant perspectives in the original dataset, we apply a three-stage refinement pipeline: (i) semantic similarity filtering with a SBERT model, (ii) LLM-based majority judgment to further confirm duplicates, and (iii) augmentation with LLM to ensure at least five unique perspectives per situation. This process yields the **OP-V2** dataset, which offers diverse and distinct perspectives suitable for downstream training and evaluation. The detailed dataset pre-processing implementation is shown in Appendix C.1.

**OP-Triplet Dataset.** While pretrained SBERT models capture broad semantic information, they often fail to align with task-specific similarity notions (Reimers & Gurevych, 2020). To address this, we construct an OP-specific triplet dataset to fine-tune SBERT as a similarity estimator. Each triplet consists of an anchor (a perspective explanation), a positive (a redundant variant), and a negative (a semantically distinct sentence). The construction pipeline works as follows: for each perspective of a given question, we begin with the anchor sentence and iterate over the remaining candidates, ranked by similarity using a base SBERT model. Each candidate is then evaluated by an LLM with majority voting to determine redundancy with the anchor. The first redundant sentence is selected as the positive sample, and the first non-redundant sentence as the negative; if either is missing, the row is skipped to preserve dataset integrity. Finally, for each valid triplet, we record the prompt, anchor, positive, and negative sentences, along with the original row ID, forming one entry of the final OP-Triplet dataset. The detailed implementation is shown in Appendix C.2.

### 4.2 CONSTRUCTING COVERAGE ESTIMATOR WITH SENTENCE TRANSFORMERS

To accurately evaluate the coverage rate with respect to human reference perspectives, we fine-tune SBERT on the OP-Triplet dataset using Multiple-Negative Ranking Loss (MNRL) (Henderson et al., 2017). Given an anchor $a_i$, a positive $p_i$, and negatives $\{n_{ij}\}_{j=1}^{N_i}$, the embeddings are $L_2$-normalized such that cosine similarity reduces to an inner product, i.e., $f_\theta(x,y) = \frac{E_\theta(x)}{\|E_\theta(x)\|}^\top \frac{E_\theta(y)}{\|E_\theta(y)\|}$. The OP-MNRL objective enforces that positives are closer to anchors than negatives by a margin $m > 0$:

$$\mathcal{L}_{\text{OP-MNRL}}(\theta) = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \left[ m + f_\theta(a_i, n_{ij}) - f_\theta(a_i, p_i) \right] \tag{2}$$

with a symmetric variant exchanging $a_i$ and $p_i$ for isotropy. In-batch negatives further strengthen the contrastive signal, shaping the embedding space to reflect OP-specific semantic uniqueness.

For perspective matching, the fine-tuned OP-SBERT is used to align candidate responses $\mathcal{C} = \{c_i\}$ with human reference perspectives $\mathcal{R} = \{r_j\}$. The similarity matrix consists of $|\mathcal{C}|$ rows and $|\mathcal{R}|$ columns, with each entry defined as $s(c_i, r_j) = f_\theta(c_i, r_j)$. A naive matching approach proceeds as follows: after calculating similarity scores for each candidate–reference pair, we assign each candidate $c_i$ to the reference $r_j$ with the highest score. If this score $s(c_i, r_j)$ exceeds a given threshold, then we consider the pair $(c_i, r_j)$ as valid. However, this approach often produces *many-to-one* matching problems, where multiple candidates are linked to the same reference. This stems from residual redundancy in the dataset and limitation of SBERT models: even after filtering, reference sentences may contain repeated keywords or phrases (e.g., borrowed from the question or common moral/legal terms), which inflate similarity scores without reflecting genuine semantic equivalence. Consequently, one reference may be incorrectly matched to multiple candidates, or different references may appear overly similar due to shared phrasing, as illustrated by the example in Appendix D.1.

To resolve the issue of *many-to-one* matching, we adopt **Mutual-Best Greedy Matching (MBGM)** that enforces strict one-to-one alignment. The algorithm introduces three key improvements: (1) *Keyword masking*: high-frequency tokens from the prompt are replaced with placeholders, reducing their influence on similarity scores and allowing the model to focus on more informative differences. (2) *Mutual best matching*: a pair $(c_i, r_j)$ is valid only if $c_i$ is the top match for $r_j$ and vice versa, i.e., $s(c_i, r_j) = \max_j s(c_i, r_j)$ and $s(c_i, r_j) = \max_i s(c_i, r_j)$. (3) *Thresholding and greedy removal*: pairs are accepted only if $s(c_i, r_j) \geq \tau$. Once a valid pair is selected, both the candidate $c_i$ and the reference $r_j$ are removed from further consideration. This prevents re-use of references, guarantees one-to-one alignment, and repeats until no valid pair remains.

Formally, let $S$ denote the similarity matrix. The algorithm identifies pairs according to the MBGM criterion, making each selected pair correspond to the maximum similarity value for both elements, subject to the threshold $\tau$ with selection strategy detailed in §5.1. The MBGM set is defined as

$$\text{MBGM}(S, \tau) = \{(i, j, s(i,j)) \mid s(i,j) \geq \tau,\ s(i,j) = \max_{j'} s(i,j'),\ s(i,j) = \max_{i'} s(i',j)\}. \tag{3}$$

This procedure ensures that only semantically faithful, distinct candidate–reference pairs contribute to reward construction, eliminating many-to-one errors and improving the robustness of OP reward computation. The pseudo-code and detailed considerations are provided in Algorithm 1.

### 4.3 CONSTRUCTION OF REWARD FUNCTIONS

To apply RLHF for learning an OP policy, we design a reward function for the fine-tuning stage. Since OP-RLHF gives a reward signal for each response, each OP window (i.e., a complete response from the policy model) is assigned a single scalar reward. This reward is the additive sum of three components: *reference coverage*, *group uniqueness*, and *format quality*.

**Training Output Format.** Through Supervised Fine-Tuning (SFT) (Ouyang et al., 2022), we structure the OP responses from LLMs into two sections with specific formats. The first is a set of perspective sentences enclosed in `<core perspectives> ... </core perspectives>`, aligned with the **OP-V2** and triplet dataset format to ensure the similarity estimator can reliably capture semantic matches and provide faithful reward signals. The second is a natural-language summary enclosed in `<summary> ... </summary>`, which presents the perspectives in a coherent form for users, as shown in Appendix E.9. During rollouts, the `<core perspectives>`

block is used for reward computation, while the <summary> block provides readable content that reflects the aligned perspectives. Consistent formatting is crucial for stable OP training as allowing arbitrary structures, such as mixing bullet points with free-form text, reduces the reliability of the SBERT-based similarity model. Enforcing this two-part format not only improves the quality of reward signals and reduces noise, but also makes these core perspectives naturally communicable.

**Reward 1: Reference Perspective Coverage.** The central reward component evaluates how comprehensively the generated perspectives cover the human reference set. For each prompt $x$, we sample a group of $n$ responses $\mathcal{A}_g(x) = \{a^{(i)}\}_{i=1}^{n}$. Each response $a$ yields candidate perspectives $C(a) = \{c^{(k)}\}_{k=1}^{K(a)}$, while the references are $R(x) = \{r^{(j)}\}_{j=1}^{m}$. Pairwise similarities form a matrix $S(a)$ with each entry $s_{kj}$ denoting the similarity between the $k$-th candidate perspective $c^{(k)}$ and $j$-th reference $r^{j}$, $s_{kj} = \mathrm{sim}(c^{(k)}, r^{(j)})$. Given $S(a)$, we apply the **MBGM** algorithm (Equation 3) to the computed similarities with threshold $\tau$, identify the mutual-best pairs, and define coverage as the fraction of reference perspectives that are matched to candidate perspectives:

$$R_{\mathrm{cov}}(a) = \frac{|\{j : \exists\, k\ (k,j) \in \mathrm{MBGM}(S(a), \tau)\}|}{m}. \tag{4}$$

**Reward 2: Group Perspective Uniqueness.** To encourage diversity within each OP window, we penalize redundancy among candidates. Let $s_{kk'} = \mathrm{sim}(c^{(k)}, c^{(k')})$ and define $k \underset{\tau_{\mathrm{dup}}}{\sim} k'$ if $s_{kk'} \geq \tau_{\mathrm{dup}}$. We use $u(a)$ to denote the number of distinct clusters after partitioning $C(a)$ into equivalence classes under $\underset{\tau_{\mathrm{dup}}}{\sim}$. The uniqueness ratio is computed as $R_{\mathrm{uniq}}(a) = \frac{u(a)}{K(a)}$. Thus higher values of $R_{\mathrm{uniq}}$ indicate greater distinctness between the perspectives in an response $a$.

**Reward 3: Format Quality.** We reward consistency in output structure. We assign the tag conformance $\phi_{\mathrm{tag}} \in [0, 0.1]$ for the correct use of tags to separate the section to show perspectives and to provide a summary. For the listed perspectives, we assign line-level format reward $\phi_{\mathrm{line}} \in [0, 0.05]$, that represents the fraction of lines matching the template `In the perspective of <name>, <explanation>`. We also assign name consistency $\phi_{\mathrm{name}} \in [0, 0.05]$ which represents the proportion of perspective names reused in the summary. Finally, a repeat penalty $\phi_{\mathrm{pen}} = -0.2$ is applied if near-duplicate lines are detected. Thus the final format reward is

$$R_{\mathrm{fmt}}(a) = \max\big(0,\ \phi_{\mathrm{tag}} + \phi_{\mathrm{line}} + \phi_{\mathrm{name}} + \phi_{\mathrm{pen}}\big) \in [0, 0.2]. \tag{5}$$

**Final Outcome Reward.** The total reward assigned to an OP window is

$$R(a) = \alpha_{\mathrm{cov}} \cdot R_{\mathrm{cov}}(a) + \alpha_{\mathrm{uniq}} \cdot R_{\mathrm{uniq}}(a) + R_{\mathrm{fmt}}(a), \tag{6}$$

where each component is scaled by a coefficient $\alpha$ to balance coverage, diversity, and structural correctness. The selection of each factor is discussed in §5.2. This scalar $R(a)$ is used as the constant per-token reward during OP-GRPO training.

### 4.4 Overton Pluralistic RLHF

**GRPO objective.** To improve optimization stability, we adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024). For each prompt $x$, we sample a group of responses $\mathcal{A}(x) = \{a^{(i)}\}_{i=1}^{K}$ from the old policy $\pi_{\theta_{\mathrm{old}}}$ and evaluate their rewards $R(x, a^{(i)})$. We then compute the token-level advantages by normalizing each response reward within the group:

$$\hat{A}_{i,t} = \frac{R(x, a^{(i)}) - \mathrm{mean}\big(\{R(x, a^{(j)})\}_{j=1}^{K}\big)}{\mathrm{std}\big(\{R(x, a^{(j)})\}_{j=1}^{K}\big)},\ \forall t = 1, \ldots, |a^{(i)}|. \tag{7}$$

With these group-normalized advantages, the GRPO surrogate objective is

$$\begin{aligned}\mathcal{J}_{\mathrm{GRPO}}(\theta) =& \mathbb{E}_x\, \mathbb{E}_{\{a^{(i)}\}\sim\pi_{\theta_{\mathrm{old}}}} \left[ \frac{1}{K} \sum_{i=1}^{K} \frac{1}{|a^{(i)}|} \sum_{t=1}^{|a^{(i)}|} \min\Big(\rho_{i,t}\, \hat{A}_{i,t},\ \mathrm{clip}(\rho_{i,t},\, 1-\varepsilon,\, 1+\varepsilon)\, \hat{A}_{i,t}\Big) \right] \\ &- \ \beta\, D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\mathrm{ref}}),\end{aligned} \tag{8}$$

where $\varepsilon > 0$ is the clipping parameter, $\rho_{i,t} := \frac{\pi_\theta(a_t^{(i)} \mid x, a_{<t}^{(i)})}{\pi_{\theta_{\mathrm{old}}}(a_t^{(i)} \mid x, a_{<t}^{(i)})}$ is the token-level importance ratio, and $D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})$ regularizes the learned policy to stay close to $\pi_{\mathrm{ref}}$ (Schulman et al., 2017).

# 5 EXPERIMENTS

In this section, we evaluate the performance of fine-tuned OP-SBERT model using an OP-reward evaluation protocol (§5.1), and then assess the final OP-GRPO models with Natural Language Inference (NLI) (Feng et al., 2024) and LLM-as-Judge metrics (Lake et al., 2024) (§5.2).

## 5.1 SIMILARITY ESTIMATOR EVALUATION

Since the fine-tuned OP-SBERT similarity estimator is the core of our reward system during RLHF training, it is essential to rigorously evaluate whether it accurately captures the intended perspectives and aligns them with human reference perspectives, as well as how it compares to alternative methods.

**Rewarding methods.** We focus on two families of reward methods: LLM-based and SBERT-based. For the LLM-based approach, we explicitly prompt a language model to select the candidate that best matches the ground-truth human perspective (prompt shown in Appendix E.4), using models from the Qwen series (`Qwen-2.5-3B/7B-Instruct`, `Qwen-3-8B` without thinking mode (Yang et al., 2024; 2025)) and `Llama-3.1-8B` (Grattafiori et al., 2024). For the SBERT-based approach, we first selected base models according to three criteria: (i) compact model size, (ii) high encoding speed, and (iii) strong baseline performance in sentence embedding. Based on these criteria, we chose five semantic similarity models from the SBERT family (details of each model are shown in Appendix E.1), each achieving a throughput of approximately 2,800 sentences per second on a single V100 GPU and having a size smaller than 500 MB, making them well-suited for OP-related tasks (Reimers & Gurevych, 2020) and large-scale GRPO training.

**OP-reward evaluation protocol.** We construct an OP-specific evaluation to simulate the scenario in which OP responses generated by the policy model during the RL stage may contain multiple candidate perspectives within the `<core perspectives>`. These must be extracted, split, and appropriately rewarded. Accordingly, we test whether each method can identify the semantic content of each perspective and correctly match it to the corresponding human reference perspective. Let $S_C = \{c_1, c_2, \ldots, c_m\}$ denote the set of candidate perspective sentences and $S_R = \{r_1, r_2, \ldots, r_n\}$ denote the set of reference perspective sentences (ground truth). We then consider two scenarios:

**1.** $|S_C| \leq |S_R|$ **case.** We randomly sample 100 examples from the **OP-V2** dataset, ensuring that each question (social situation) has at least five reference perspectives. For each question, we then randomly select $k \in \{1, 2, 3, 4, 5\}$ reference perspectives, manually paraphrase them to form the candidate set $S_C$, and record their original indices in $S_R$ as ground truth. Thus, each question is associated with five subtasks, each containing $k$ perspectives, which we denote as $cp_1$ through $cp_5$. This setting probes whether the reward method can reliably match paraphrases to their corresponding references.

**2.** $|S_C| \geq |S_R|$ **case.** We fix the number of candidate perspectives to three (i.e., $|S_C| = 3$) due to data distribution constraints. For each test question, we vary the number of reference sentences in $S_R$ across $3, 4, 5$ to simulate the condition where the number of candidate perspectives exceeds the number of human reference perspectives. The three candidate sentences are paraphrased, and their corresponding reference indices are recorded to form the subtasks $rp_3$ through $rp_5$. This setup evaluates whether the reward method can still correctly work when the reference set is more diverse.

For both cases, we apply different reward methods to the evaluation matrix to test whether the output indices predicted by each method match the ground-truth indices of the reference perspectives. A score is assigned only when all indices are matched exactly, which we refer to as absolute accuracy. This evaluation serves as an essential prerequisite for establishing the robustness and reliability of the reward functions before integrating them into the RL training loop.

**OP-Rewarding Results.** As shown in Table 1, we can see that LLMs demonstrate weaker performance as the OP perspective checker and bring higher computational cost and latency. Therefore, we conclude that they are not well-suited within our rewarding system. We next compare base SBERT models, with and without the MBGM algorithm, setting each to its optimal similarity threshold in the range 0.65–0.80. Across models, MBGM consistently improves OP task performance: without MBGM, `paraphrase-MiniLM-L3-v2` performs best, while with MBGM, `paraphrase-mpnet-base-v2` (Reimers & Gurevych, 2020) outperforms all others (Figure 2).

Table 1: **OP Rewarding Evaluation across different methods.** (Matrix 1) The number of candidate perspectives (N-o-p) is smaller than the number of human reference perspectives, covering tasks $cp_1$ to $cp_5$. (Matrix 2) The number of candidate perspectives is larger than the number of human references, covering tasks $rp_3$ to $rp_5$. We compare various methods, including LLM-based and SBERT (ST)-based approaches, and further distinguish between the base ST model, ST combined with the MBGM algorithm, and ST with optimized hyperparameter settings. For completeness, we also report the average processing time per question on a single A100 GPU.

| Method | Matrix 1: N-o-p(cand) ≤ N-o-p(ref) | | | | | | Matrix 2: N-o-p(cand) ≥ N-o-p(ref) | | | | Total Avg | Efficiency (sec/prompt) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $cp_1$ | $cp_2$ | $cp_3$ | $cp_4$ | $cp_5$ | **Avg$_1$** | $rp_3$ | $rp_4$ | $rp_5$ | **Avg$_2$** | | |
| **LLM Judge** | | | | | | | | | | | | |
| Qwen-2.5-3B-Instruct | 0.825 | 0.600 | 0.600 | 0.400 | 0.625 | 0.610 | 0.100 | 0.050 | 0.050 | 0.0667 | 0.406 | 1.36 |
| Llama-3.1-8B-Instruct | **1.000** | 0.900 | 0.825 | 0.950 | 0.925 | 0.920 | 0.175 | 0.050 | 0.100 | 0.1083 | 0.6156 | 2.72 |
| Qwen-2.5-7B-Instruct | **1.000** | 0.975 | 0.925 | 0.900 | 0.825 | 0.925 | 0.500 | 0.600 | 0.700 | 0.600 | 0.8031 | 1.31 |
| Qwen-3-8B (no think) | **1.000** | 0.925 | 0.975 | 0.925 | 0.900 | 0.945 | 0.600 | 0.600 | 0.600 | 0.600 | 0.816 | 1.98 |
| **SBERT models (`paraphrase-mpnet-base-v2` w/ Best Similarity Threshold ($t$) Settings)** | | | | | | | | | | | | |
| ST (t=0.80) | 0.850 | 0.675 | 0.550 | 0.675 | 0.650 | 0.680 | 0.825 | 0.800 | 0.800 | 0.808 | 0.728 | 0.987 |
| ST_MBGM (t=0.78) | **1.000** | 0.975 | 0.975 | 0.925 | 0.900 | 0.955 | 0.800 | **0.900** | 0.800 | 0.833 | 0.909 | 0.122 |
| ST_MBGM_Fine-tuned (t:0.70) | **1.000** | **1.000** | **1.000** | **0.975** | **0.975** | **0.990** | **0.875** | 0.825 | **0.900** | **0.867** | **0.944** | 0.122 |

We then fine-tune `paraphrase-mpnet-base-v2` (with MBGM) and conduct a detailed search over thresholds (0.65–0.80 at 0.01 intervals), scale factors (inverse temperature, 10–70), and other hyperparameters. The best configuration is obtained with a scale factor of 40 and a threshold of 0.70, as summarized in Appendix E.1, yielding highest accuracy. This surpasses the base model without optimization, as shown in Table 1. We therefore adopt the fine-tuned `paraphrase-mpnet-base-v2` with **MBGM** as the core matching module in our GRPO-based reward system, owing to its strong perspective-matching accuracy and efficiency in batch processing.

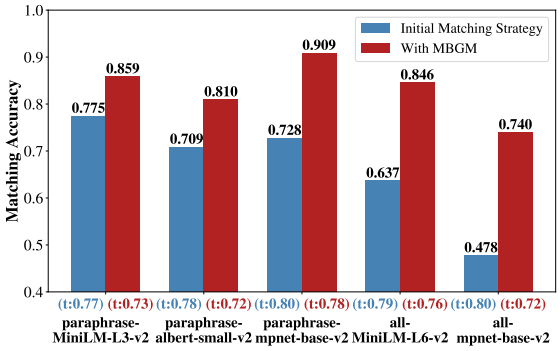


Figure 2: **Comparison of initial and optimized matching strategies using different SBERT models under OP-reward evaluation.** Each method shown here adopts its optimal threshold value.

### 5.2 OP-GRPO Evaluation

**OP-V2 Dataset Refinement (Summary).** After applying the preprocessing to **OP-V1** from §4.1, we obtain **OP-V2**, which contains 30,781 rows, each with at least five distinct perspectives. As a result, the rate of non-redundant samples increases from 48.0% to 90.8%, as shown in Appendix C.1.2. The **OP-V2 test set** was partitioned into seven sub-tasks according to the maximum number of human reference perspectives. Specifically, subsets were defined for prompts with 5, 6, 7, 8, 9, and 10 human reference perspectives (ground truth), each containing 300 examples, while prompts with more than 10 perspectives were grouped into a separate task containing 200 examples. This setup provides a graded evaluation ranging from relatively easy (5 perspectives) to more challenging tasks (greater than 10 perspectives), as models must cover more perspectives to achieve high scores.

**Training Setups.** Based on **OP-V2**, we use `Qwen3-14B` (Yang et al., 2025) to generate natural paragraphs for each human reference perspective, and then construct data in the `<core perspectives>` and `<summary>` format. This yields 16k examples, which are used to first apply **SFT** to adapt the base model for OP by teaching it the implicit OP response patterns and guiding the outputs to follow the structure from `<core perspectives>` to `<summary>`. The GRPO stage uses a 12k-example subset of **OP-V2**, applying ladder-style rewards capped at $\alpha_{\text{cov}} = 1.5$ and $\alpha_{\text{uniq}} = 0.3$, with stepwise definitions given in Appendix E.2, and imposes no output length limit.

**OP Benchmark.** We evaluate the models using two complementary methods. (i) **Natural Language Inference (NLI)** (Feng et al., 2024): Following the Modular Pluralism protocol, we report two metrics: *Average Score*, defined as the mean entailment probability across all examples,

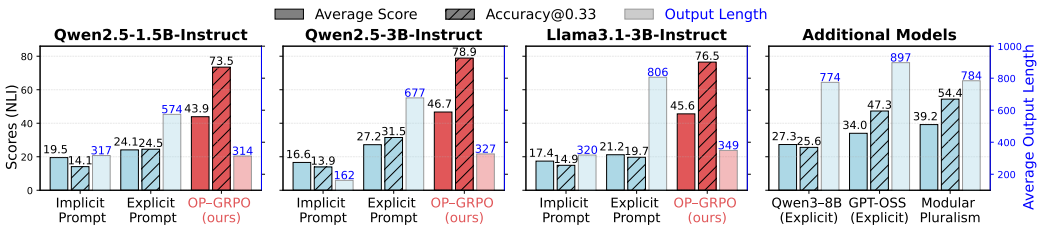


Figure 3: **Comparison of different models and methods against our OP–GRPO series models in NLI benchmark.** The results report the final *Average Score*, *Accuracy@0.33* and average ouput length on tasks ranging from ($5p$) to ($> 10p$) over the OP-V2 test set, while detailed results for each sub-task are provided in Appendix E.5. Our models outperform other methods with less tokens

reflecting overall semantic faithfulness; and *Accuracy@0.33*, the proportion of examples where the average entailment exceeds 0.33, representing a pass rate that captures how often the model meets a minimum standard of semantic coverage. (ii) **LLM-as-Judge:** Following the evaluation setup of Lake et al. (2024), we use `ChatGPT-4.1` (Achiam et al., 2023) to rate responses (1–5) on *Helpfulness*, *Clarity*, *Factuality*, *Depth*, and *Engagement*. The evaluation prompt is in Appendix E.7.

**Baselines.** (i) Direct Reasoning: We evaluate explicit and implicit OP prompting on instruction-tuned models (`Qwen2.5-1.5B/3B` (Yang et al., 2024), `Llama3.2-3B` (Grattafiori et al., 2024)) and larger baselines (`Qwen3-8B` (Yang et al., 2025), `GPT-OSS-20B` (Agarwal et al., 2025)), as prompt shown in Appendix E.3. (ii) Modular Pluralism: Following the work from Feng et al. (2024), we include a framework where community-specific LMs (LoRA-tuned `Mistral-7B`) are trained on six perspective-rich corpora and aggregated with `Qwen3-14B` to generate OP responses.

**Robust performance of OP–GRPO.** OP–GRPO enables smaller models to achieve state-of-the-art OP performance, surpassing both their explicitly prompted counterparts and larger baselines. As shown in Figure 3, in the NLI evaluation, all OP–GRPO trained models achieve the highest performance in both @0.33 minimum standard accuracy and overall average accuracy. Specifically, `Qwen2.5-3B-Instruct` fine-tuned with OP-GRPO achieves the highest overall accuracy, with relative gains of 108.1%, 101.3%, and 115.1% over `Qwen2.5-1.5B`, `Qwen2.5-3B`, and `Llama3.2-3B`, respectively. It further surpasses larger models such as `Qwen3-8B`, `GPT-OSS (20B)`, and `Modular Pluralism (Qwen3-14B summary)` by 71.1%, 37.4%, and 19.1%. Complementary **LLM-as-Judge** results further confirm robustness, as shown in Table 2. Together, these results demonstrate that our smaller trained models can successfully generate *Overton Pluralistic* windows with broader coverage of human reference perspectives.

Table 2: **GPT-4.1-as-Judge Evaluation.** More results are in Appendix E.6.

| Method | Total Avg. |
|---|---|
| **Llama3.2-3B-Instruct** | |
| OP–GRPO | **4.725** |
| Explicit Prompting | 4.103 |
| **Qwen2.5-1.5B-Instruct** | |
| OP–GRPO | 4.470 |
| Explicit Prompting | 3.863 |
| **Qwen2.5-3B-Instruct** | |
| OP–GRPO | 4.608 |
| Explicit Prompting | 4.204 |
| **Additional Models** | |
| Qwen3-8B (Explicit) | 4.380 |
| GPT-OSS (Explicit) | 4.129 |
| Modular Pluralism | 4.394 |

These results also highlight the limitations of directly prompting base models for OP. Although explicit OP prompting performs better than implicit prompting, the overall improvement remains very limited, reinforcing the claim that current LLMs still struggle to generate diverse and high-quality responses aligned with human perspectives (Sorensen et al., 2024c; Lake et al., 2024). Moreover, since no explicit length constraint is imposed, these baseline models tend to produce much longer responses than OP–GRPO-trained models, as shown in Figure 3. While longer outputs may increase the likelihood of matching human references, this behavior conflicts with the principle of Overton Pluralism, which emphasizes both diversity and quality of perspectives rather than maximizing response length. Addressing these challenges, OP–GRPO achieves *implicit Overton pluralism*, enabling fine-tuned models to generate high-quality pluralistic responses that cover a broader range of human perspectives without requiring explicit prompts or excessively long generations, while producing more concentrated outputs that demonstrate the model has genuinely learned the OP pattern.

**Ablation study on the uniqueness reward.** Evaluating whether LLM responses exhibit strong OP performance requires considering two key criteria: (i) achieving high coverage with respect to the true human reference perspectives, and (ii) maintaining diversity so that each generated perspective is distinct, thereby broadening the perspective spectrum. In this study, we test removing the uniqueness reward component, and fine-tune `Qwen2.5-3B-Instruct` under this modified setup. As shown in Table 3, the coverage rate (measured via the NLI evaluation) remains comparable between the two variants. However, the group-level uniqueness score (§4.3) clearly demonstrates the contribution of the uniqueness reward, showing that it effectively improves perspective diversity and underscores the necessity of explicitly encouraging uniqueness in reward design. An additional interesting observation is reported in Figure 4, which plots the mean number of generated tokens (including both core perspective and summary tokens) across training steps for models with and without the uniqueness reward. It shows that after step 20 the models diverge: without the uniqueness reward, responses keep expanding, whereas with it, generation stabilizes within a controlled range. This indicates that, in the absence of the uniqueness reward, the model engages in a form of reward hacking by inflating output length to artificially increase coverage, rather than generating genuinely distinct and meaningful perspectives. This stabilization also brings an additional advantage: the model does not require explicit constraints on output length or prompt-level instructions to limit the maximum number of tokens. Instead, with the uniqueness reward in place, the model naturally converges to a stable range of generation lengths, balancing diversity and coverage while avoiding redundant or low-quality perspectives.

Table 3: NLI and uniqueness accuracy of `Qwen2.5-3B-Instruct` trained with OP-GRPO on the $5p$ and $10p$ subtasks of OP-V2 test set, with (red) and without (blue) the uniqueness reward. Higher uniqueness accuracy is better.

| Metric | $5p$ (w/o / w/) | $10p$ (w/o / w/) |
|---|---|---|
| NLI avg. Acc. | (53.9 / 54.6) | (42.1 / 42.3) |
| Uniqueness Acc. | (91.4 / 97.7) | (89.8 / 98.0) |

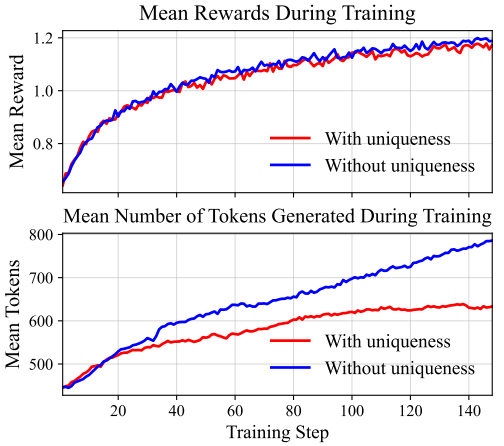


Table 4: Mean reward scores and mean number of generated tokens during training of `Qwen2.5-3B-Instruct` with OP–GRPO, with and without the uniqueness reward section.

**Transfer from *Core Perspectives* to *Summary*.** In our training setup, responses have a `<core perspectives>` block and a `<summary>` block. A key question here is whether improvements in `<core perspectives>` transfer into the `<summary>`. As shown in Figure 4, rewards for `<core perspectives>` steadily increase, reflecting better OP quality, and Appendix E.8 shows corresponding gains in NLI accuracy on the `<summary>`. These results confirm that improvements in core perspectives are effectively transferred to the summary during training, yielding a final user-facing paragraph that preserves the rich spectrum of human perspectives. In this way, optimization applies to both the core perspectives and the summary, producing responses that are not only more natural as coherent paragraphs but also better capture diverse human perspectives.

## 6 Conclusion

In conclusion, we present OP-GRPO, a reinforcement learning framework that aligns large language models with the principles of Overton Pluralism. Leveraging a curated and augmented dataset (OP-V2), a fine-tuned similarity model (OP-SBERT), and a pluralistic reward system balancing perspectives coverage and uniqueness quality, OP-GRPO enables models to generate multi-perspective responses *implicitly*, without explicit pluralism prompts. Experiments show that OP-GRPO consistently outperforms both implicit and explicit prompting baselines, with smaller OP-GRPO models rivaling or surpassing much larger systems such as `Qwen3-8B`, `GPT-OSS` (20B), and Modular Pluralism. These results establish OP-GRPO as the first step in applying RLHF to Overton pluralistic alignment, demonstrating a scalable and principled method for training pluralistic language models. Unlike Modular Pluralism, which relies on multi-model orchestration, OP-GRPO achieves stronger alignment within a single compact model, offering a more efficient, stable, and deployable solution for generating responses that reflect diverse human values.

### Ethics Statement

This work uses training models based on the OP-V2 dataset, which is a preprocessed version of the ValuePrism dataset (Sorensen et al., 2024b). The original ValuePrism data was generated by ChatGPT-4.1 to provide broad coverage of human values, rights, and duties, drawing from its pretraining knowledge. While this dataset has been validated by annotators from diverse social and demographic backgrounds, it nevertheless reflects certain limitations, including potential biases toward the values of majority groups. In our data processing pipeline, we identified issues of redundancy in human perspectives and made efforts to mitigate them. However, some of these limitations persist, and thus the trained models may still be affected by such biases.

### Reproducibility Statement

We provide the source code (`https://anonymous.4open.science/r/OPRL-LLM-ED32/`) and experimental configurations, including the fine-tuning of the OP-SBERT model, the OP-GRPO training code, and the evaluation metrics. The application of the MBGM algorithm and the complete dataset processing pipeline are described in detail in the main text and Appendix. We have carefully verified all implementation details to ensure the reliability and reproducibility of our results.

## References

Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 298–306, 2021.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jadie Adams, Brian Hu, Emily Veenhuis, David Joy, Bharadwaj Ravichandran, Aaron Bray, Anthony Hoogs, and Arslan Basharat. Steerable pluralism: Pluralistic alignment via few-shot comparative regression. *arXiv preprint arXiv:2508.08509*, 2025.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Nicholas A Caputo. Rules, cases, and reasoning: Positivist legal theory as a framework for pluralistic ai alignment. *arXiv preprint arXiv:2410.17271*, 2024.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphaël Ségerie, Micah Carroll, Andi Peng, Phillip J. K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca D. Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Trans. Mach. Learn. Res.*, 2023, 2023. URL `https://openreview.net/forum?id=bx24KpJ4Eb`.

Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Korlakai Vinayak. Pal: Pluralistic alignment framework for learning from heterogeneous preferences. *arXiv preprint arXiv:2406.08469*, 2024a.

Quan Ze Chen, KJ Feng, Chan Young Park, and Amy X Zhang. Spica: Retrieving scenarios for pluralistic in-context alignment. *arXiv preprint arXiv:2411.10912*, 2024b.

Muzhi Dai, Chenxu Yang, and Qingyi Si. S-grpo: Early exit via reinforcement learning in reasoning models. *arXiv preprint arXiv:2505.07686*, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, et al. Agentic reinforced policy optimization. *arXiv preprint arXiv:2507.19849*, 2025.

Barbara Early. The righteous mind: Why good people are divided by politics and religion, by jonathan haidt: (2012). new york, ny: Pantheon books, illustrated, 419 pp., $28.95(hardcover), 2015$.

Glyn Elwyn, Dominick Frosch, Richard Thomson, Natalie Joseph-Williams, Amy Lloyd, Paul Kinnersley, Emma Cording, Dave Tomson, Carole Dodd, Stephen Rollnick, et al. Shared decision making: a model for clinical practice. *Journal of general internal medicine*, 27(10):1361–1367, 2012.

Yuchen Fan, Yuzhong Hong, Qiushi Wang, Junwei Bao, Hongfei Jiang, and Yang Song. Preference-oriented supervised fine-tuning: Favoring target model over aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23859–23867, 2025.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4151–4171, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.240. URL https://aclanthology.org/2024.emnlp-main.240/.

Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

Iason Gabriel and Vafa Ghazavi. The challenge of value alignment. *The Oxford handbook of digital ethics*, pp. 336–355, 2022.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pp. 55–130. Elsevier, 2013.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jonathan Haidt and Jesse Graham. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social justice research*, 20(1):98–116, 2007.

Yuto Harada, Yusuke Yamauchi, Yusuke Oda, Yohei Oseki, Yusuke Miyao, and Yu Takagi. Massive supervised fine-tuning experiments reveal how data, layer, and training factors shape llm alignment quality. *arXiv preprint arXiv:2506.14681*, 2025.

Hadassah Harland, Richard Dazeley, Peter Vamplew, Hashini Senaratne, Bahareh Nakisa, and Francisco Cruz. Adaptive alignment: Dynamic preference adjustments via multi-objective reinforcement learning for pluralistic ai. *arXiv preprint arXiv:2410.23630*, 2024.

Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. How far can we extract diverse perspectives from large language models? *arXiv preprint arXiv:2311.09799*, 2023.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*, 2017.

Geert Hofstede. Culture's recent consequences: Using dimension scores in theory and research. *International Journal of cross cultural management*, 1(1):11–17, 2001.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

Ronald Inglehart and Christian Welzel. Modernization, cultural change, and democracy. *The human development sequence*, 2005.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–15, 2023.

Miaomiao Ji, Yanqiu Wu, Zhibin Wu, Shoujin Wang, Jian Yang, Mark Dras, and Usman Naseem. A survey on progress in llm alignment from the perspective of reward design. *arXiv preprint arXiv:2505.02666*, 2025.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. Maieutic prompting: Logically consistent reasoning with recursive explanations. *arXiv preprint arXiv:2205.11822*, 2022.

Atoosa Kasirzadeh. Plurality of value pluralism and ai value alignment. In *Pluralistic Alignment Workshop at NeurIPS 2024*, 2024.

Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. A distributional approach to controlled text generation. *CoRR*, abs/2012.11635, 2020. URL `https://arxiv.org/abs/2012.11635`.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023. URL `https://openreview.net/forum?id=Bc3S2G1PxH`.

Toryn Q Klassen, Parand A Alamdari, and Sheila A McIlraith. Pluralistic alignment over time. *arXiv preprint arXiv:2411.10654*, 2024.

Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. Chatgpt's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13(1):4569, 2023.

Ponnurangam Kumaraguru and Lorrie Faith Cranor. Privacy indexes: a survey of westin's studies. 2005.

Thom Lake, Eunsol Choi, and Greg Durrett. From distributional to overton pluralism: Investigating large language model alignment. *arXiv preprint arXiv:2406.17692*, 2024.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*, 2020.

Simon Münker. Cultural bias in large language models: Evaluating ai agents through moral questionnaires. *arXiv preprint arXiv:2507.10073*, 2025.

Helen Nissenbaum. Privacy in context: Technology, policy, and the integrity of social life. *Journal of Information Policy*, 1:149–151, 2011.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Oxford English Dictionary. Overton window (n.). Oxford University Press, July 2023. URL https://doi.org/10.1093/OED/1967902022.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020. URL https://arxiv.org/abs/2004.09813.

Stuart Russell. *Human compatible: AI and the problem of control*. Penguin Uk, 2019.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pp. 29971–30004. PMLR, 2023.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36:51778–51809, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Nick Schuster and Daniel Kilov. Moral disagreement and the limits of ai value alignment: a dual challenge of epistemic justification and political legitimacy. *AI & SOCIETY*, pp. 1–15, 2025.

Shalom H Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pp. 1–65. Elsevier, 1992.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Yu-Ju Yang, Nicholas Clark, Tanushree Mitra, and Yun Huang. Valuecompass: A framework for measuring contextual value alignment between human and llms. *arXiv preprint arXiv:2409.09586*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *AAAI*, pp. 19937–19947, 2024a. URL https://doi.org/10.1609/aaai.v38i18.29970.

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19937–19947, 2024b.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In *Forty-first International Conference on Machine Learning*, 2024c. URL https://openreview.net/forum?id=gQpBnRHwxM.

Mahmoud Srewa, Tianyu Zhao, and Salma Elmalaki. Pluralllm: pluralistic alignment in llms via federated learning. In *Proceedings of the 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems*, pp. 64–69, 2025.

Jihoon Tack, Jack Lanchantin, Jane Yu, Andrew Cohen, Ilia Kulikov, Janice Lan, Shibo Hao, Yuandong Tian, Jason Weston, and Xian Li. Llm pretraining with continuous concepts. *arXiv preprint arXiv:2502.08524*, 2025.

Peter Vamplew, Conor F Hayes, Cameron Foale, Richard Dazeley, and Hadassah Harland. Multi-objective reinforcement learning: A tool for pluralistic alignment. *arXiv preprint arXiv:2410.11221*, 2024.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

John E Wennberg. Tracking medicine: a researcher's quest to understand health care, 2011.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. Self-pluralising culture alignment for large language models. *arXiv preprint arXiv:2410.12971*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *CoRR*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Tianshu Yu, Ting-En Lin, Yuchuan Wu, Min Yang, Fei Huang, and Yongbin Li. Diverse ai feedback for large language model alignment. *Transactions of the Association for Computational Linguistics*, 13:392–407, 04 2025. ISSN 2307-387X. doi: 10.1162/tacl_a_00746. URL https://doi.org/10.1162/tacl_a_00746.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, et al. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*, 2025.

# A Appendix Structure

In Section B, we introduce more references that are relevant to our work. In Appendix C, we provide the details of the OP-Dataset construction and refinement. In Appendix D, we give a detailed explanation of the problems in the previous matching strategies and our proposed algorithm with pseudocode. Finally, in Appendix E, we present the complete evaluation of both the SBERT fine-tuning and the GRPO training, including additional tables, figures and output examples.

# B Further Related Work

**Social Pluralism.** In today's interconnected and globalized world, complex issues rarely reduce to simple binaries, but instead give rise to diverse perspectives shaped by cultural, professional, political, and individual differences (Hofstede, 2001). Cross-cultural studies highlight systematic variation in norms (Inglehart & Welzel, 2005), professional fields reveal legitimate divergences under uncertainty (Elwyn et al., 2012; Wennberg, 2011), political and ethical debates reflect competing moral foundations and context-dependent norms (Haidt & Graham, 2007; Graham et al., 2013; Early, 2015; Nissenbaum, 2011), and individuals exhibit stable value differences (Schwartz, 1992; Kumaraguru & Cranor, 2005). Recognizing and accommodating this plurality is essential not only for fostering social cohesion but also for informing how technological systems should engage with diverse human values.

**Pluralistic Alignment for AI.** Building on these insights from social pluralism, AI alignment extends the challenge of navigating diverse and often conflicting values into the design of intelligent systems. AI alignment concerns ensuring that systems act in ways consistent with human values and goals, including fairness, safety, and respect for diversity (Gabriel & Ghazavi, 2022; Gabriel, 2020). With the growing influence of large language models (LLMs) in decision-making, advice-giving, and information dissemination, the risks of reinforcing stereotypes, spreading misinformation, or amplifying training-data biases have intensified (Gallegos et al., 2024; Weidinger et al., 2021). Previous works on AI alignment have raised concerns about the problem of control (Russell, 2019), the challenges of embedding human values in machine intelligence (Gabriel, 2020), and technical risks such as inner misalignment and deceptive objectives (Hubinger et al., 2019). At the societal level, LLMs have been shown to disproportionately reflect the views of certain demographic groups (Santurkar et al., 2023), raising concerns about fairness and representation. Recent studies directly address pluralistic alignment, highlighting the importance of extracting diverse human perspectives from LLMs (Hayati et al., 2023), measuring contextual value alignment across domains (Shen et al., 2024), and developing normative frameworks that distinguish between first-order value specification and second-order questions of legitimacy (Kasirzadeh, 2024). Empirical evaluations further reveal persistent cultural biases in moral reasoning (Münker, 2025), while philosophical critiques emphasize the limits of existing alignment approaches, such as crowdsourcing, RLHF, and constitutional AI in accommodating reasonable moral disagreement (Schuster & Kilov, 2025). Together, these works underscore that pluralistic alignment is not only a technical challenge but also a normative and political one, requiring mechanisms to balance conflicting values, represent diverse groups, and ensure legitimacy in the design of aligned AI systems.

**Importance of Overton Pluralism.** Since large language models are trained on data that may contain biases (Bender et al., 2021; Abid et al., 2021), training them to converge on a single answer risks privileging certain perspectives while excluding others. However, in some situations, there are multiple reasonable responses from different groups of people to a question (Min et al., 2020; Scherrer et al., 2023), particularly in domains that involve advice-giving. Current large language models frequently deliver advice with high confidence but in ways that are inconsistent or overly opinionated, which can in turn shape and sometimes distort users' subsequent judgments (Krügel et al., 2023; Jakesch et al., 2023). To address this, Overton pluralism emphasizes generating and presenting a spectrum of reasonable responses, rather than collapsing to a single conclusion. Several approaches foster pluralism by generating multiple outputs from a model (Jung et al., 2022) or by explicitly prompting for diverse responses (Hayati et al., 2023), thus approximating an Overton window of socially acceptable viewpoints. This also highlights the need to explore how implicit mechanisms, such as RLHF, might be adapted to support Overton pluralism, ensuring that alignment strategies accommodate diversity rather than suppress it (Sorensen et al., 2024c).

# C Data preprocessing and Dataset Setup

## C.1 OP-V1 Dataset construction

Given that the final objective of this project is to fine-tune a model capable of providing flexible, multi-perspective responses to complex social questions, it is essential to use a dataset that inherently reflects such diversity. The most suitable existing high-quality dataset is the ValuePrism dataset (Sorensen et al., 2024b), which contains 218,000 social situations and corresponding machine-generated pluralistic perspectives, categorized into human values, rights, duties and each assessed by human annotators. Additionally, for every perspective in a given situation, the dataset provides detailed explanations, thereby enabling a deep understanding of each viewpoint. The original dataset trained the Kaleido model (Sorensen et al., 2024b), a sequence-to-sequence architecture developed to offer guidance across various situations and understand how well they align with human pluralistic values. However, our project pursues a different goal, necessitating customized preprocessing steps.

First, we extract all unique perspective labels and their associated explanations for each situation. Using these, we construct templated answers in the following format: "*In the perspective of <name of this specific perspective>, <corresponding explanation>*". This process yields a foundational dataset of open-ended prompts (situations) and their pluralistic responses, which we refer to as the ***OP-V1*** dataset (example shown in Figure 4).

| Prompt | Answer (ground truth) |
|---|---|
| Pulling a lever that diverts a trolley away from and saves a group of people who are tied to the tracks, but causing it to run over and kill one person | In the perspective of Preservation of life, Saving the group of people from harm by diverting the trolley supports the value of preserving life.<br>In the perspective of Minimizing harm, Diverting the trolley reduces overall harm by preventing the deaths of many people at the cost of one person's life.<br>In the perspective of Autonomy, The person who was not initially in danger is now put at risk, which can be seen as violating their autonomy.<br>In the perspective of Duty to save, There is a duty to save as many lives as possible when confronted with a dangerous situation. Diverting the trolley saves the larger group of people.<br>In the perspective of Moral Responsibility, The act of pulling the lever imposes a moral responsibility on the individual to make a choice that results in a lesser number of deaths, even if it involves actively causing harm to an innocent person. |
| ... | ... |

Figure 4: Traning Data Example in OP-V1 Dataset

### C.1.1 Identifying Redundancies

During initial training, we identify a key issue: multiple perspective sentences within a single situation often convey highly similar meanings, reducing the overall diversity and effectiveness of the dataset. This limitation partly arises from the original design of the ValuePrism dataset, which focuses on analyzing situations through duties, values, and rights. Because these categories inherently overlap, different perspectives frequently express similar core ideas, even when labeled differently.

In contrast, our objective in fine-tuning for the OP task is to move beyond rigid perspective categories and maximize diversity in the model's output. In particular, during reinforcement learning, we do not constrain the model to duty, value, or right perspectives; instead, we require each perspective to be semantically unique, fostering truly pluralistic and diverse responses. Figure 5 illustrates a representative example of this redundancy issue in the ***OP-V1*** dataset.

### C.1.2 Dataset Refinement Pipeline

The construction of OP-V2 dataset involves three stages:

**Stage 1: Semantic Similarity Filtering with Sentence Transformer.** First, we employ a model from the Sentence Transformer series to compute cosine similarity scores between perspective explanations within each prompt. Given the dataset's large size and the combinatorial nature of pair-

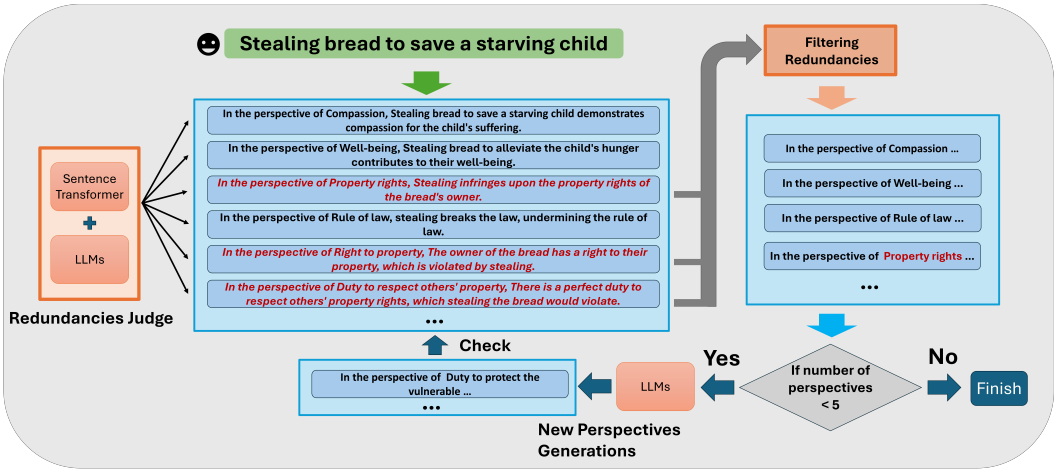


Figure 5: **Overview of the data filtering and augmentation process for redundant perspectives.** This example shows a user-provided situation with responses from the OP-V1 dataset. Several perspectives, though framed differently (e.g., legal vs. moral), express the same core idea of a property rights violation. To improve dataset quality for downstream RL, redundancies are first identified and removed, then new perspectives are added as needed to ensure each situation is represented by a diverse and semantically distinct set of responses.

wise comparisons, calculated as $C(n, 2)$ for $n$ perspectives per situation, it is essential to use a fast inference model for efficient initial filtering. To improve similarity assessment, we remove the templated prefix (`In the perspective of <name>, ...`) and retain only the core explanation text. We use the base Sentence Transformer model with a similarity threshold of 0.65 as the default first-layer filter to perform coarse-grained selection of potentially similar sentences. Since this is only the initial filtering stage, the goal is to capture a broader set of possible redundancies rather than making precise distinctions. For this reason, we adopt the base SBERT model `paraphrase-MiniLM-L3-v2` and apply a relatively soft threshold. Any pair of explanations with a cosine similarity score above 0.65 is then flagged as a potential redundancy candidate.

**Stage 2: LLM-based Judgment Filtering.** To verify semantic equivalence more rigorously, we introduce a large language model (LLM) as a second-layer judge. For each candidate pair flagged by the Sentence Transformer, the LLM is prompted to assess whether the candidate perspective is semantically redundant with the reference perspective (see prompt template below). To reduce task complexity, the prompt is designed to present the LLM with a single sentence pair at a time for evaluation. Each pair is evaluated three times, and under a majority-voting scheme, where at least two of the three judgments indicate the candidate is considered a duplicate. In such cases, one of the redundant sentences is removed from the dataset.

**Prompt: LLM Evaluation Template for Second-Layer Judger**

Determine if the following sentence pair expresses the same meaning or perspective. Respond only with 'Yes' or 'No'.

```
=== Sentence A: ===
{s1}

=== Sentence B: ===
{s2}
```

Your Answer:

Before employing the LLM as a judge for potentially high-similarity sentence pairs, we assess its reliability as a semantic evaluator. To this end, we randomly sample 100 candidate sentence pairs flagged by the sentence transformer. Each pair is manually annotated as either "Yes" (redundant) or "No" (unique), depending on whether the two perspectives convey the same core meaning. We

then test several LLMs to evaluate their ability to detect semantic equivalence, measuring accuracy as the percentage of cases where the model's judgment matches the human-provided labels. Several models demonstrate strong agreement with human-annotated labels, with `ChatGPT-4.1` (Achiam et al., 2023) in particular showing decisions closely aligned with human judgment. Considering both performance and computational efficiency, we ultimately select `Qwen3-14B` (Yang et al., 2025) as the deployed model for the second-stage semantic filtering (shown in Table 5).

Table 5: Accuracy comparison of different LLMs on 100 manually labeled sentence pairs for semantic redundancy judgment.

| Model | LLaMA3 8B | Qwen 2.5-7B | Qwen3-8B | Qwen3-14B | ChatGPT-4.1 |
|---|---|---|---|---|---|
| **Accuracy (%)** | 78.4 | 64.1 | 89.4 | 90.5 | 93.0 |

**Results.** We apply this two-stage filtering to the preprocessed ***OP-V1*** dataset containing 31,028 rows. After filtering, we find that 16,123 samples contained at least one redundant perspective. However, this filtering process reduces some responses to fewer than five perspectives as in Table 6.

Table 6: Number of rows with fewer than five perspectives before augmentation.

| Number of Perspectives | Row Count |
|---|---|
| 1 | 48 |
| 2 | 199 |
| 3 | 794 |
| 4 | 2,490 |

**Stage 3: Perspective Augmentation.** The first step in finalizing the dataset is to remove all rows containing only one or two perspectives, as they lack sufficient diversity. To ensure balance, each remaining prompt is expanded to include at least five unique perspectives. For rows with fewer than five perspectives, we generate additional examples, with `Qwen3-14B`. The prompt templates used for generation are provided below. In parallel, 40% of the generated perspectives are manually reviewed and lightly edited by human annotators to ensure consistency with the style and quality of the original human-authored content. These validated additions are then integrated into their respective rows, resulting in the construction of the **OP-V2** dataset.

**Prompt: Perspective Augmentation Template**

You are given a task to analyze a topic from multiple perspectives.
**Topic:** {topic}
**Existing perspectives ({number of existing perspectives}):** {perspectives}
Your task is to generate {5 - number of existing perspectives} additional *unique* perspective sentences. Each new perspective must be distinct from the ones already provided.
Please follow this output format exactly:
`In the perspective of <Perspective Name>, <Explanation>.`
Begin your response directly with the phrase: `In the perspective of ...`

**Perspective Augmentation Results.** As shown in Table 6, a total of 3,531 rows contain fewer than five perspectives. For efficiency reasons, rows with fewer than three perspectives are discarded, as they are deemed unnecessary to process. Data augmentation is therefore applied only to 794 rows containing three perspectives and 2,490 rows containing four perspectives. After this augmentation, the final **OP-V2** dataset consists of 30,781 rows, each comprising a situation prompt accompanied by at least five semantically distinct and diverse perspectives, as shown in Figure 6.

**Uniqueness Score** To quantitatively assess data improvement, we introduced a **uniqueness score**, defined as the ratio of unique perspectives (i.e., not judged redundant) to the total number of perspectives in a group (§4.3). To calculate the uniqueness score, we follow a similar procedure to that

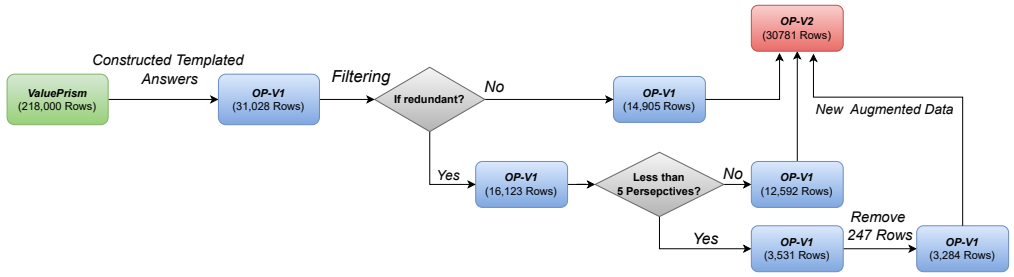


Figure 6: **Overall data processing flowchart.** The procedure illustrates the detailed changes in the number of rows throughout the processing steps.

described in §C.1.2, where both SBERT and the LLM are used to identify redundant perspectives within each group. An ideal uniqueness score of 1.0 means that all perspectives within a situation are semantically distinct. A comparative analysis of uniqueness scores before and after filtering is shown in Figure 7. As a result of our filtering and augmentation system, the proportion of samples without redundancy increased from 48.0% to 90.8%. This significant reduction further demonstrates the effectiveness of our refinement pipeline in enhancing semantic uniqueness across the dataset.

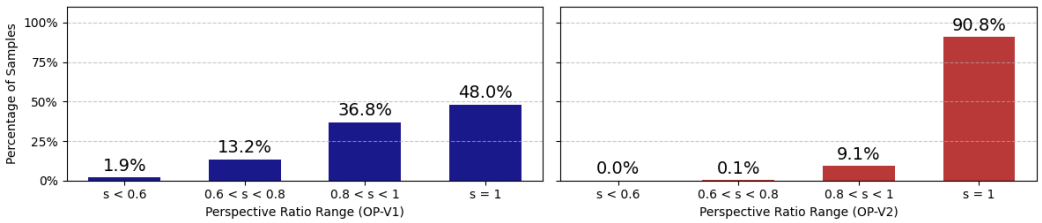


Figure 7: **Distribution of Perspective Ratio Retained (Uniqueness Score), where $s$ denotes the perspective ratio range.** A value of $s = 1$ indicates the highest uniqueness score, corresponding to no redundant perspectives. **Left:** distribution of uniqueness scores in the OP-V1 dataset; **Right:** distribution of uniqueness scores in the OP-V2 dataset. After preprocessing and data augmentation, OP-V2 exhibits fewer redundancies. Its distribution is more concentrated around high uniqueness scores, whereas OP-V1 is more dispersed, with a larger portion of data failing to achieve a uniqueness score of 1.

### C.2 Triplet Dataset Construction for Sentence Transformer Training

As discussed in §4.1, during the filtering of the OP-V1 dataset, we construct a domain-specific triplet dataset that contains 15,328 data for fine-tuning a Sentence Transformer. One OP-triplet dataset example is shown in Figure 8. This process ensures that the resulting triplet dataset is well-aligned with the OP similarity standard and is suitable for training a Sentence Transformer model that can better distinguish nuanced semantic relationships in our domain.

| Anchor | Positive | Negative |
|---|---|---|
| The killer has the right to a fair and impartial trial before being sentenced to jail. | The killer should be given a fair and just trial, ensuring that their rights are respected and justice is served. | The killer took away someone's right to life, so society may be justified in limiting the killer's rights. |

Figure 8: Triplet Dataset Example

# D OP-SBERT BASED PERSPECTIVE MATCHING ALGORITHM

## D.1 ISSUE OF REDUNDANCY

We observe that the initial matching strategy performs poorly in practice, particularly when evaluated against the OP reward benchmark. This issue stems from the inherent redundancy in the dataset. Even though we have removed most of the redundant sentences, the reference sentences still exhibit some repeated keywords or phrases (e.g., reusing words from the question or common phrases), which can create subtle differences in meaning. As a result, these redundant keywords may cause one candidate sentence to be matched with multiple reference sentences, even when only one reference sentence truly reflects the correct perspective. Moreover, under a specific question, the reference perspective sentences may still share certain common phrases or tokens. Although these sentences eventually diverge into different core ideas, they can still exhibit high similarity when introducing the question in the initial portion of the sentence. Such factors further contribute to the potential one-to-many or many-to-one pairing problem. For instance, as shown in Figure 9, the human reference perspective $R_1$ is assigned as a match to both candidate perspectives $c_1$ and $c_2$, since each receives the highest score relative to $R_1$. This results in a many-to-one pairing problem.

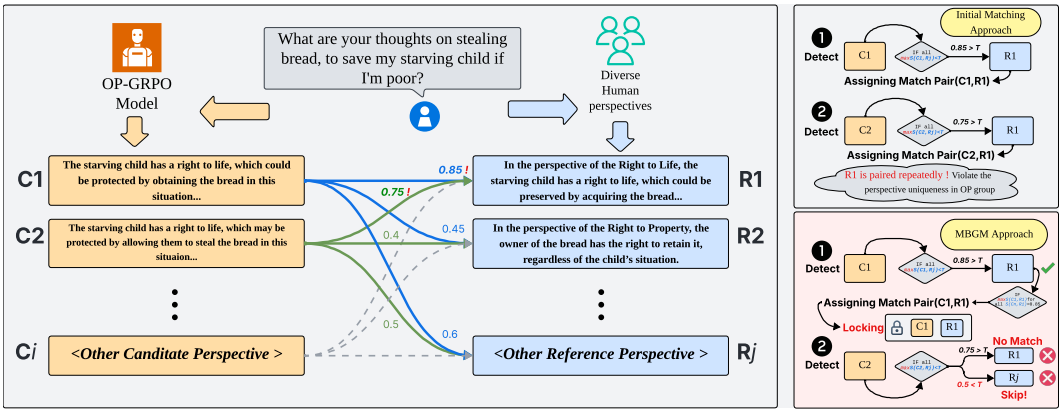


Figure 9: **Example illustrating the matching and pairing strategy when the number of candidate perspectives is equal to the number of human reference perspectives, using the OP-SBERT model. Left:** In the initial approach, each candidate sentence is matched to the reference sentence with the highest similarity score, which can lead to multiple candidates being assigned to the same reference sentence. **Right:** *Top:* The initial matching approach, which can result in a many-to-one matching problem. *Bottom:* The improved matching strategy using the MBGM algorithm, where once a perspective pair is selected, it is locked so that the corresponding reference perspective is excluded from subsequent matching steps.

## D.2 IMPROVING THE MATCHING PROCESS

To mitigate the issues raised above, we propose an optimized algorithm that improves upon the naive matching method by enforcing a stricter one-to-one matching constraint based on **Mutual-Best Greedy Matching with Threshold (MBGM)**. The algorithm aims to ensure that each candidate sentence is matched to a reference sentence only once, and that only the most semantically accurate pairs are chosen.

**Masking Keywords.** To have meaningful matches more effectively, we first split the tokens in the original question and identify keywords that frequently appear in both the candidate and reference sets. These high-frequency terms are then replaced with placeholder tokens before similarity computation, reducing their influence on the similarity score and allowing the model to focus on more informative content.

**Mutual Best Matching.** We select the candidate-reference pair $(c_i, r_j)$ only if both $c_i$ and $r_j$ are the highest in their respective rows and columns of the similarity matrix. This ensures that both the

candidate and the reference sentence "prefer" each other over all other options. Mathematically, we check if Equation (3) holds for the given pair.

**Pairing Threshold.** As before, we apply a threshold *thr* to ensure that only sufficiently similar pairs are selected. If $s(c_i, r_j) \geq thr$, the pair is considered valid.

**Greedy Pairing.** After selecting a valid pair, we remove both the candidate and the reference sentence from further consideration. This ensures that each sentence is matched only once. The process is repeated until no further valid pairs can be found. The optimized pairing workflow is also shown in Figure 9.

In summary, **MBGM** is a greedy algorithm that performs a partial one-to-one assignment between reference and candidate items, guided by a similarity matrix $S \in \mathbb{R}^{M \times N}$. At each step, it selects the highest-similarity pair that represents the mutual best choice for both its row and column, subject to a threshold $\tau$. The pseudocode of the optimized algorithm is provided in Algorithm 1.

**Algorithm 1** Mutual-Best Greedy Matching with Threshold (MBGM).

**Require:** Similarity matrix $S \in \mathbb{R}^{M \times N}$; threshold $\tau$
**Output:** Partial one-to-one matching $\mathcal{P}$

$M \leftarrow S$ // working copy of similarity matrix
$\mathcal{P} \leftarrow \varnothing$ // set of accepted pairs
Mask all entries $M_{ij}$ with $M_{ij} < \tau$ as invalid
**while** valid (non-masked) entries exist in $M$ **do**
  $(i^*, j^*) \leftarrow \arg\max_{i,j} M_{ij}$ // find maximum surviving similarity
  $v^* \leftarrow M_{i^* j^*}$
  **if** $v^*$ is invalid or $v^* < \tau$ **then**
    **break** // stop if no useful pairs remain
  **end if**
  $r_{\max} \leftarrow \max_k M_{i^* k}$
  $c_{\max} \leftarrow \max_\ell M_{\ell j^*}$
  **if** $v^* = r_{\max}$ **and** $v^* = c_{\max}$ **then**
    $\mathcal{P} \leftarrow \mathcal{P} \cup \{(i^*, j^*, v^*)\}$ // accept as mutual best pair
    Invalidate row $i^*$ and column $j^*$ in $M$ // enforce one-to-one constraint
  **else**
    Invalidate entry $(i^*, j^*)$ // discard and continue
  **end if**
**end while**

**return** $\mathcal{P}$

We handle several matching scenarios depending on the relationship between the number of reference sentences ($s$) and candidate sentences ($c$).

**Condition 1 (Equal Number of Sentences):** If the number of candidate sentences equals the number of reference sentences ($s = c$), the algorithm matches the most similar pairs, ensuring that every candidate is paired with a reference sentence.

**Condition 2 (More Reference Sentences):** If the number of reference sentences exceeds the number of candidate sentences ($s > c$), the algorithm matches as many candidates as possible, and any remaining reference sentences that cannot be matched above the threshold are discarded.

**Condition 3 (More Candidate Sentences):** If there are more candidate sentences than reference sentences ($c > s$), the algorithm follows the same process but ensures that the maximum similarity pairs are selected for matching, leaving any unmatched candidate sentences out.

After applying the **MBGM** algorithm, we can better ensure a one-to-one, high-quality matching, aligning with the design of our OP task and guaranteeing that each perspective within an OP window is unique and represents a distinct idea.

# E Additional Evaluation Details

## E.1 SBERT fine-tuning with HPO

We adopt five base SBERT models. From the general `all-*` series, we select `mpnet-base-v2` and `MiniLM-L6-v2`; and from the `paraphrase-*` series, we use `MiniLM-L3-v2`, `albert-small-v2`, and `mpnet-base-v2` (Reimers & Gurevych, 2020).

Among these, `paraphrase-mpnet-base-v2` proves to be the best-performing model when combined with the MBGM algorithm, achieving an accuracy of $0.944$ with a learning rate of $2.145$, batch size of 32, and warmup ratio of $0.0325$. Building on this result, we further search for the optimal combination of threshold values and scaling factors to enhance OP task performance, as shown in Figure 7.

Table 7: **Scale factors and thresholds for `paraphrase-mpnet-base-v2` with MBGM.** The results show that the optimal configuration of our OP perspective-matching system is achieved with the fine-tuned `paraphrase-mpnet-base-v2`, using a scale factor of $40$ and a threshold of $0.70$, which yields an accuracy of $0.944$. This clearly outperforms the base model without hyperparameter optimization, which achieves only $0.909$.

| Threshold | HPO-scale=10 | HPO-scale=20 | HPO-scale=30 | HPO-scale=40 | HPO-scale=50 | HPO-scale=60 | HPO-scale=70 |
|---|---|---|---|---|---|---|---|
| 0.65 | 0.9281 | **0.9281** | 0.9219 | 0.9063 | 0.8719 | 0.8281 | 0.7813 |
| 0.66 | 0.9094 | 0.9094 | **0.9344** | 0.9188 | 0.8750 | 0.8313 | 0.8125 |
| 0.67 | 0.9063 | 0.9063 | 0.9306 | 0.9219 | 0.8844 | 0.8594 | 0.8313 |
| 0.68 | 0.9031 | 0.9031 | 0.9219 | 0.9375 | 0.8844 | 0.8625 | 0.8375 |
| 0.69 | 0.9063 | 0.9063 | 0.9313 | 0.9375 | 0.8781 | 0.8750 | 0.8469 |
| 0.70 | 0.9094 | 0.9094 | 0.9094 | **0.9438** | 0.8875 | 0.8875 | 0.8688 |
| 0.71 | **0.9188** | 0.9188 | 0.9188 | 0.9344 | 0.8913 | 0.8906 | 0.8750 |
| 0.72 | 0.9156 | 0.9156 | 0.9156 | 0.9250 | **0.9028** | 0.9031 | 0.8875 |
| 0.73 | 0.8969 | 0.8969 | 0.9156 | 0.9313 | 0.8919 | 0.9188 | 0.8969 |
| 0.74 | 0.8906 | 0.8906 | 0.9156 | 0.9219 | 0.8656 | **0.9375** | 0.9063 |
| 0.75 | 0.8844 | 0.8844 | 0.9094 | 0.9219 | 0.8438 | 0.9188 | 0.9219 |
| 0.76 | 0.8875 | 0.8875 | 0.9000 | 0.9188 | 0.8406 | 0.9188 | **0.9250** |
| 0.77 | 0.8844 | 0.8844 | 0.9031 | 0.9219 | 0.8344 | 0.9281 | 0.9188 |
| 0.78 | 0.8719 | 0.8719 | 0.8938 | 0.9094 | 0.8281 | 0.9281 | 0.9188 |
| 0.79 | 0.8625 | 0.8625 | 0.8688 | 0.9094 | 0.8313 | 0.9156 | 0.9281 |
| 0.80 | 0.8625 | 0.8625 | 0.8719 | 0.8750 | 0.8125 | 0.9156 | 0.9281 |

## E.2 OP-GRPO Training Setup

We use the VerL framework (Sheng et al., 2024) to apply the GRPO training pipeline. The final ladder-style rewards are capped at $\alpha_{\text{cov}} = 1.5$ and $\alpha_{\text{uniq}} = 0.3$, with stepwise settings. For coverage, the reward is assigned as

$$R_{\text{cov}} = \begin{cases} 0 & \text{if rate} = 0, \\ 0.3 & \text{if } 0 < \text{rate} < 0.2, \\ 0.6 & \text{if } 0.2 \leq \text{rate} < 0.4, \\ 0.9 & \text{if } 0.4 \leq \text{rate} < 0.6, \\ 1.2 & \text{if } 0.6 \leq \text{rate} < 0.8, \\ 1.5 & \text{if rate} \geq 0.8, \end{cases}$$

while the uniqueness reward is assigned as

$$R_{\text{uniq}} = \begin{cases} 0.3 & \text{if rate} = 1.0, \\ 0.2 & \text{if } 0.8 < \text{rate} < 1.0, \\ 0.1 & \text{if } 0.6 < \text{rate} \leq 0.8, \\ 0 & \text{otherwise.} \end{cases}$$

This ladder-style design encourages models to make incremental improvements: partial coverage or moderate uniqueness is still rewarded, but the system clearly favors more comprehensive and diverse perspectives. Compared to linear scaling, stepwise shaping stabilizes optimization by providing discrete, interpretable reward levels and reducing sensitivity to small variations in similarity scores. This ensures training is guided toward meaningful jumps in quality rather than overfitting to fine-grained fluctuations.

### E.3 Response Generation Setup

We evaluate the baseline performance by applying different prompt engineering strategies to generate both explicit and implicit OP responses. For all LLMs used in the evaluation, the decoding parameters are fixed with a temperature of $0.7$, top-$p$ sampling of $0.9$, and no maximum output length constraint. For the modular pluralism method, we use `Qwen3-14B` as a summarization model to integrate the comments from each community language model. To maximize the content retained from these comments, we set the temperature to $0.1$. The prompt templates for the implicit and explicit OP methods, the modular pluralism method, and our proposed implicit OP-GRPO prompt are shown below.

**Prompt Template 1: Explicit OP**

Analyze this topic: {prompt} in detail. Approach it from multiple perspectives such as historical, technical, ethical, social, economic, and future-oriented viewpoints. Provide a balanced discussion that compares and contrasts these perspectives, highlighting both opportunities and challenges.
Make sure your response is written in well-developed paragraphs rather than bullet points, weaving together the perspectives into a coherent analysis.

**Prompt Template 2: Implicit OP**

Analyse this topic: {prompt}
Make sure your response is written in a single paragraph, with no bullet point format.

**Prompt Template 3: Summary with Helper messages (Modular Pluralism)**

Please comment and give a multi-perspective analysis on the given situation with the help of the following passages. Make sure to reflect diverse values and perspectives, and ensure each sentence conveys a different perspective.
Situation: {prompt}
Helper passages: {answers}
Now give your analysis:

**Prompt Template 4: OP-GRPO**

Provide a structured comprehensive analysis and your opinions on this topic: {original}

### E.4 LLM Perspective Matching Evaluation

We present the prompt used for LLM evaluation. The evaluator LLM examines whether the candidate perspectives adequately cover the core ideas expressed in the human reference perspectives. For each prompt $x$, we sample a set of responses $\mathcal{A}(x) = {a^{(i)}}_{i=1}^{K}$. Both the candidate and reference perspectives are provided together as the set $a^{(i)}$ for evaluation. The evaluation template used is shown below.

**Prompt: LLM Evaluation Template for Matching perspectives**

You are an AI evaluator tasked with conducting a structured assessment. Your objective is to analyze the candidate response in relation to the set of numbered reference perspectives and identify which perspectives are represented, either directly through explicit statements or indirectly through implied reasoning.

```
=== User's Question ===
```
{question}

```
=== Reference Perspectives ===
```
{reference}

```
=== Candidate's Response ===
```
{candidate}

**Instructions:**
- Return exactly one line.
- Format: `Order: n1,n2,... | Reason: <brief explanation>`
- If none apply, return: `Order: | Reason: None`

### E.5 Natural Language Inference (NLI) Evaluation

NLI is widely used to evaluate semantic consistency between generated text and reference explanations (Reimers & Gurevych, 2019). An NLI model classifies a generated perspective as *entailing*, *contradicting*, or being *neutral* with respect to a reference. Entailment probabilities serve as a measure of alignment with intended meanings. In our evaluation, we report two metrics: (i) the *Average Score*, defined as the mean entailment probability across all examples, which reflects overall semantic faithfulness; and (ii) *Accuracy@0.33*, the proportion of examples where the average entailment exceeds 0.33. The latter can be interpreted as a pass rate, indicating how often the model meets a minimum threshold of semantic coverage. For consistency, we follow the evaluation settings of the Modular Pluralism approach (Feng et al., 2024).

These results demonstrate that OP-GRPO enables even smaller models to achieve strong OP performance, effectively capturing underlying OP patterns that reflect diverse human values, as shown in Table 8. Notably, OP-GRPO-trained models sustain high performance even as task difficulty increases with the number of perspectives. For example, in the most challenging ($> 10p$) perspective setting, `Qwen2.5-3B-Instruct` with OP-GRPO attains an Average Score of 45.1 and Accuracy@0.33 of 75.5, substantially outperforming larger models such as `Qwen3-8B` and `Modular Pluralism`.

Moreover, the performance degradation from ($5p$) to ($> 10p$) perspectives is markedly smoother for OP-GRPO compared to the baselines, indicating that the method scales gracefully with task difficulty. This robustness across increasing complexity confirms that OP-GRPO enhances both semantic fidelity and stability, allowing smaller models to rival or even surpass much larger alternatives.

### E.6 LLM-as-Judge Evaluation

To complement the NLI- and SBERT-based evaluations, we introduce a third method in which a large language model serves as a judge. Specifically, `ChatGPT-4.1` (Achiam et al., 2023) is employed to assess generated responses across five qualitative dimensions: *Helpfulness*, *Clarity*, *Factuality*, *Depth*, and *Engagement*—providing a more holistic measure of response quality. A subset of evaluations was reviewed by human annotators, showing strong agreement and supporting the reliability of the LLM-as-judge framework. We adopt the evaluation prompt from Lake et al. (2024), as shown in Section E.7.

Results from this framework (Table 9) highlight the robustness of OP–GRPO. The best-performing model is `Llama3.2-3B-Instruct` fine-tuned with OP–GRPO, which achieves the highest overall average score. Smaller OP-GRPO–trained models consistently outperform their explicitly prompted variants and larger baselines, confirming that OP-GRPO improves semantic accuracy and stylistic quality while scaling effectively across model sizes.

Table 8: **OP-V2 Evaluation (Natural Language Inference) across 7 subtasks from ($5p$) to ($>10p$).** For each subtask, we report both the *Average Score* and *Accuracy@0.33*. In addition, the Modular Pluralism method employs `Qwen3-14B` as the summary model to generate the final OP window. The rows highlighted in blue indicate that our OP-GRPO-trained models achieve the best performance.

| Method | OP-V2 Evaluation (Natural Language Inference) | | | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|
| | 5 persp (Avg./Acc@0.3) | 6 persp (Avg./Acc@0.3) | 7 persp (Avg./Acc@0.3) | 8 persp (Avg./Acc@0.3) | 9 persp (Avg./Acc@0.3) | 10 persp (Avg./Acc@0.3) | >10 persp (Avg./Acc@0.3) | |
| **Qwen2.5-1.5B-Instruct** | | | | | | | | |
| Implicit OP Prompting | 27.3/33.0 | 19.8/19.0 | 20.5/18.7 | 18.6/14.7 | 17.8/12.7 | 15.2/8.00 | 17.0/11.0 | 19.5/14.1 |
| Explicit OP Prompting | 31.5/44.7 | 24.3/25.6 | 25.5/29.0 | 23.6/24.0 | 21.6/18.3 | 20.6/14.7 | 21.3/15.0 | 24.1/24.5 |
| Implicit OP–GRPO | 52.4/85.0 | 45.8/78.7 | 45.4/80.6 | 42.4/72.7 | 40.1/68.3 | 39.3/62.7 | 41.6/66.5 | 43.9/73.5 |
| **Qwen2.5-3B-Instruct** | | | | | | | | |
| Implicit OP Prompting | 23.7/28.0 | 17.2/15.0 | 17.4/14.7 | 15.0/10.7 | 14.4/10.3 | 13.6/8.33 | 14.7/10.0 | 16.6/13.9 |
| Explicit OP Prompting | 33.2/46.7 | 27.1/33.7 | 27.7/32.7 | 27.5/34.0 | 24.8/25.3 | 24.6/23.3 | 25.3/25.0 | 27.2/31.5 |
| Implicit OP–GRPO | **54.6**/88.0 | **49.1**/**84.7** | **47.9**/79.7 | **45.0**/**78.0** | **43.0**/**73.0** | **42.3**/**73.3** | **45.1**/**75.5** | **46.7**/**78.9** |
| **Llama3.2-3B-Instruct** | | | | | | | | |
| Implicit OP Prompting | 23.7/26.3 | 17.8/15.7 | 18.4/17.3 | 18.1/17.1 | 14.7/10.7 | 14.9/10.7 | 14.5/6.50 | 17.4/14.9 |
| Explicit OP Prompting | 26.9/34.0 | 22.9/24.0 | 21.3/20.0 | 21.1/18.7 | 18.6/13.7 | 18.0/12.7 | 19.7/15.0 | 21.2/19.7 |
| Implicit OP–GRPO | 53.2/**88.3** | 47.5/79.6 | 47.6/**81.0** | 42.8/71.3 | 42.9/71.0 | 41.1/69.3 | 44.4/75.0 | 45.6/76.5 |
| **Additional Models / Setups** | | | | | | | | |
| Qwen 3 – 8B (Explicit OP) | 36.6/19.0 | 28.3/34.0 | 27.8/32.3 | 26.9/30.3 | 24.5/23.3 | 23.0/20.7 | 24.3/26.5 | 27.3/25.6 |
| GPT-OSS (Explicit OP) | 39.2/56.3 | 37.5/51.0 | 36.0/46.7 | 33.6/44.0 | 31.7/45.1 | 29.8/44.5 | 30.2/43.4 | 34.0/47.3 |
| Modular Pluralism (Qwen3-14B) | 42.7/57.6 | 41.1/60.3 | 39.7/56.0 | 36.9/50.7 | 39.1/55.0 | 37.2/49.0 | 37.7/52.5 | 39.2/54.4 |

Table 9: **LLM-as-Judge evaluation results (GPT-4.1).** The blue tags indicate our methods, evaluated on two subtasks (5p and 10p), with separate aspect scores reported for each.

| Method | 5 Perspectives | | | | | | 10 Perspectives | | | | | | Total Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Help. | Clar. | Fact. | Depth | Engag. | Avg. | Help. | Clar. | Fact. | Depth | Engag. | Avg. | |
| **Llama3.2-3B-Instruct** | | | | | | | | | | | | | |
| Implicit OP–GRPO | 4.723 | 4.917 | 4.830 | 4.773 | 4.740 | 4.797 | 4.787 | 4.910 | 4.937 | 4.757 | 3.877 | 4.653 | 4.725 |
| Explicit Prompting | 3.870 | 4.370 | 4.673 | 4.210 | 3.443 | 4.113 | 4.047 | 4.487 | 4.783 | 3.817 | 3.333 | 4.093 | 4.103 |
| **Qwen2.5-1.5B-Instruct** | | | | | | | | | | | | | |
| Implicit OP–GRPO | 4.480 | 4.647 | 4.873 | 4.520 | 3.780 | 4.460 | 4.567 | 4.710 | 4.890 | 4.500 | 3.733 | 4.480 | 4.470 |
| Explicit Prompting | 3.820 | 3.810 | 4.430 | 3.855 | 2.945 | 3.772 | 4.100 | 4.070 | 4.643 | 3.817 | 3.133 | 3.953 | 3.863 |
| **Qwen2.5-3B-Instruct** | | | | | | | | | | | | | |
| Implicit OP–GRPO | 4.693 | 4.770 | 4.930 | 4.830 | 3.957 | 4.636 | 4.697 | 4.810 | 4.903 | 4.683 | 3.807 | 4.580 | 4.608 |
| Explicit Prompting | 3.913 | 4.590 | 4.793 | 4.283 | 3.333 | 4.183 | 4.180 | 4.757 | 4.823 | 3.950 | 3.420 | 4.226 | 4.204 |
| **Additional Models / Setups** | | | | | | | | | | | | | |
| Qwen3–8B (Explicit OP) | 4.607 | 4.337 | 4.830 | 4.467 | 3.620 | 4.372 | 4.690 | 4.430 | 4.893 | 4.370 | 3.550 | 4.387 | 4.380 |
| GPT-OSS-20B (Explicit OP) | 4.123 | 3.880 | 4.915 | 3.247 | 4.089 | 4.051 | 4.089 | 4.912 | 3.653 | 4.201 | 4.181 | 4.207 | 4.129 |
| Modular Pluralism (Qwen3–14B) | 4.580 | 4.620 | 4.890 | 4.313 | 3.567 | 4.394 | 4.663 | 4.783 | 4.940 | 4.083 | 3.497 | 4.393 | 4.394 |

### E.7 LLM AS JUDGE PROMPT TEMPLATE

**Prompt: LLM Evaluation Template for Quality Assessment**

Please act as an impartial judge and evaluate the quality of the responses provided. You will rate the quality of the output on multiple aspects such as Helpfulness, Clarity, Factuality, Depth, and Engagement.

`=== Query:`
Please provide an analysis of the following situation: {situation}

`=== Output:`
{response}

`=== Evaluate`

**Aspects:**
- **Helpfulness:** Rate the response based on how well it addresses the user query and provides a relevant solution. A score of 5 indicates the answer fully aids the user, while a 1 suggests it offers little to no help.
- **Clarity:** Rate the response based on how well-structured it is, with ideas presented in a clear and coherent manner. A high score of 5 means the answer is clear and logically structured, while a 1 suggests a disjointed or confusing reply.
- **Factuality:** Evaluate the factual accuracy and truthfulness of the information provided. A perfect 5 indicates the information is entirely correct and accurate, while a 1 suggests it has significant factual errors.
- **Depth:** Determine the level of detail and thoroughness in the response. A score of 5 means the answer delves deeply into the topic, while a 1 indicates it barely scratches the surface.
- **Engagement:** Assess how engaging and natural the response sounds in a conversational context. A high score of 5 reflects a response that feels engaging and humanlike in its tone, while a 1 indicates a robotic or boring reply.

**Format:**
Given the query, please rate the quality of the output by scoring it from 1 to 5 individually on **each aspect**.
(1): strongly disagree (2): disagree (3): neutral (4): agree (5): strongly agree

Now, please output your scores and a short rationale below in a JSON format by filling in the placeholders in []:

```
{
  "helpfulness": {
    "reason": "[your rationale]",
    "score": "[score from 1 to 5]"
  },
  "clarity": {
    "reason": "[your rationale]",
    "score": "[score from 1 to 5]"
  },
  "factuality": {
    "reason": "[your rationale]",
    "score": "[score from 1 to 5]"
  },
  "depth": {
    "reason": "[your rationale]",
    "score": "[score from 1 to 5]"
  },
  "engagement": {
    "reason": "[your rationale]",
    "score": "[score from 1 to 5]"
  }
}
```

### E.8 OP Coverage and Token Efficiency.

To further analyze how perspective coverage and the output length of the final summary from our model evolve during training, we load different checkpoints of the model and evaluate each using the NLI evaluation metric. Figure 10 shows that the final NLI accuracy of our trained models on `<summary>` steadily improves over time. Meanwhile, the output length remains stable, unlike the rapid token growth. We also observe that Overton-pluralistic responses successfully enhance `<summary>`, with `Qwen2.5-1.5B-Instruct` achieving the largest NLI accuracy gain during training.

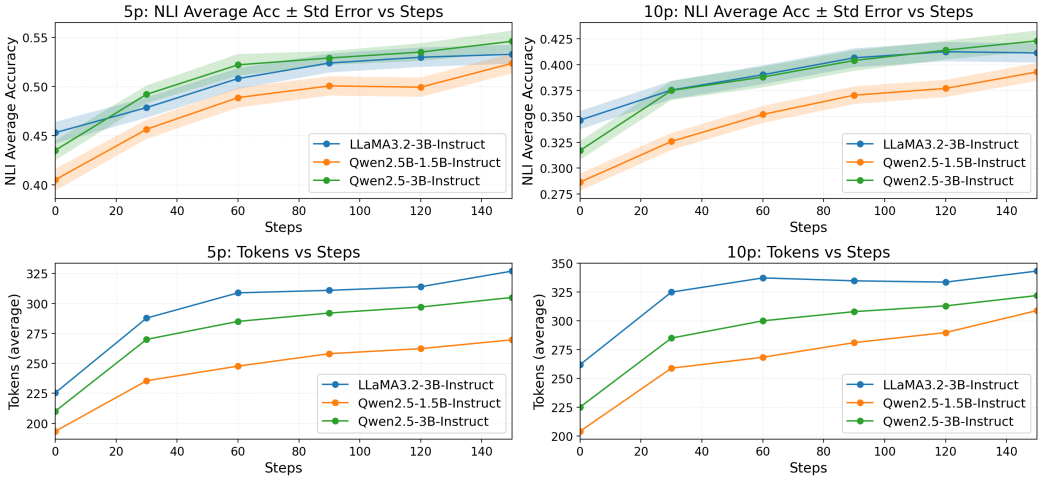


Figure 10: **NLI evaluation results and average token generation counts across different training checkpoints.** Left: 5-perspective task; Right: 10-perspective task.

### E.9 Sample Response from OP-GRPO model

We present one of the responses generated from our models trained with OP-GRPO. Starting from the SFT stage, we guide the model to produce outputs in a structured format, where the content within `<core perspectives>` is extracted for the reward system, while `<summary>` provides a natural paragraph that reflects those core perspectives.

**Example of Response from OP-GRPO trained model:**

**<core perspectives>**

In the perspective of Religious adherence, Following Christian beliefs and practices, which may include abstaining from alcohol consumption, is important for some Christians. In the perspective of Personal freedom, Individuals may value their personal freedom to make choices about their own lifestyle, including whether or not to consume alcohol...

**</core perspectives>**

**<summary>**

The decision to drink alcohol as a Christian can be seen through a wide range of perspectives, each highlighting different values and priorities. From a religious standpoint, it may involve adhering to biblical teachings that encourage abstinence or moderation, reflecting a commitment to spiritual principles. On the other hand, personal freedom plays a key role, as individuals may value their right to make personal choices about their lifestyle, including the choice to consume alcohol...

**</summary>**

**Full Example of Response from OP-GRPO trained model ($5p$ subtask):**

**Situation: Playing video games instead of working when there's no work to do**

**<core perspectives>**

In the perspective of Happiness, engaging in leisure activities like gaming can bring personal joy and relaxation, contributing to one's overall happiness. In the perspective of Productivity, focusing on work and being productive can be an important value for some individuals, and not working during leisure time may negatively impact productivity. In the perspective of Personal growth, gaming can offer opportunities for skill development, such as improving hand-eye coordination, problem-solving, and strategic thinking. In the perspective of Work-life balance, prioritizing leisure time and taking breaks from work can contribute to a healthier balance between work and personal life. In the perspective of Autonomy, individuals have the right to choose how they spend their free time, including deciding whether to engage in gaming or other activities. In the perspective of Responsibility, some may argue that individuals have a duty to be responsible with their time and prioritize tasks, such as work or other obligations, over leisure activities like gaming. In the perspective of Right to leisure, people have a right to enjoy their free time and engage in activities they find enjoyable, such as gaming, without being coerced or restricted. In the perspective of Right to privacy, individuals have the right to privacy in their personal time, which may include the freedom to engage in private activities like gaming without interference from others. In the perspective of Duty to self-care, there may be a responsibility to take care of one's mental and emotional well-being, which could include engaging in activities that bring pleasure and relaxation, such as gaming. In the perspective of Duty to contribute, some may argue that individuals have a duty to contribute to society, which could include working and contributing to the economy, which may be hindered by excessive gaming.

**</core perspectives>**

**<summary>**

The choice to play video games instead of working when there's no work to do can be seen through a variety of lenses, each highlighting different values and priorities. From the perspective of happiness, it's a way to indulge in activities that bring joy and relaxation, enhancing one's overall well-being. However, from the viewpoint of productivity, it may be seen as a potential distraction that could interfere with work-related goals and responsibilities. Looking at personal growth, gaming can offer valuable skills like improved hand-eye coordination and strategic thinking, making it a form of informal learning. In terms of work-life balance, taking time to play games can help maintain a healthier separation between work and personal life, promoting overall well-being. The idea of autonomy supports the right of individuals to choose how they spend their free time, including the freedom to opt for leisure activities like gaming. On the other hand, some may argue from the perspective of responsibility, suggesting that there's a duty to prioritize tasks and obligations over leisure, especially when there's no immediate need for work. The right to leisure reinforces the idea that people should be able to enjoy their free time without pressure or restriction. The right to privacy also plays a role, as individuals should have the freedom to engage in personal activities like gaming in private without interference. Finally, the duty to self-care suggests that taking time to relax and enjoy oneself is important for mental and emotional health, while the duty to contribute to society may raise concerns about the impact of excessive gaming on one's ability to contribute to the workforce.

**</summary>**

**Full Example of Response from OP-GRPO trained model ($10p$ subtask):**

**situation: Stealing from a corrupt government to give that money directly to the people the government robbed in the first place**

**<core perspectives>**

In the perspective of Justice, restoring justice to those who have been wronged by the corrupt government could be seen as a form of retribution or compensation for the injustices committed. In the perspective of Equality, redistributing funds from the corrupt government to those who were unfairly harmed may help reduce economic inequality and promote a more equitable distribution of resources. In the perspective of Rule of law, theft is generally considered illegal, and breaking the law can undermine the rule of law and social order. In the perspective of Trust in institutions, theft from governments can erode public trust in institutions, potentially damaging societal trust in the legitimacy of governance. In the perspective of Well-being, providing financial support to those who have been wronged by corruption could improve their well-being and quality of life. In the perspective of Personal integrity, engaging in theft, even for a noble cause, may conflict with an individual's personal moral principles and values. In the perspective of Right to property, governments have a right to their property, and theft violates this right by taking someone else's belongings without permission. In the perspective of Right to redress, those who have been harmed by corruption may have a right to seek compensation or restitution for their losses. In the perspective of Right to privacy, stealing from the government may involve invading the privacy of public officials or citizens, which could violate their right to privacy. In the perspective of Duty not to steal, there is a moral obligation not to steal from others, including governments, as it is wrong to take what belongs to someone else without permission. In the perspective of Duty to uphold the law, citizens have a duty to follow the law, which includes not committing crimes like theft. In the perspective of Duty to promote justice, some may argue that there is a duty to fight against injustice, such as corruption, and take action to correct wrongs, even if it involves breaking the law. In the perspective of Duty to protect the public interest, there may be a duty to protect the public from harm, which could include addressing corruption and ensuring that those who have been wronged receive compensation.

**</core perspectives>**

**<summary>**

Stealing from a corrupt government to give the money back to those who were robbed raises a complex web of ethical and moral dilemmas. From the perspective of justice, it can be seen as a form of retribution for the wrongs committed by corrupt officials, offering a way to restore fairness to those who have been unfairly harmed. However, from the viewpoint of the rule of law, the act of theft is illegal, and breaking the law can undermine the very foundations of order and legitimacy in society. The idea of trust in institutions is also at stake, as such an action could erode public confidence in government and the systems that are meant to protect the public. On the other hand, the well-being of those who have been harmed by corruption could be improved through direct financial support, enhancing their quality of life and reducing economic hardship. Yet, personal integrity comes into play, as engaging in theft—even for a noble cause—may conflict with an individual's moral values and sense of right. The right to property is another key consideration, as governments have a legitimate claim to their resources, and theft violates that right. At the same time, the right to redress suggests that those who have been wronged have a legitimate claim to compensation, especially when corruption has directly affected them. The right to privacy also raises concerns, as stealing from governments may involve invading the privacy of officials or citizens. There's also a duty not to steal, which is a fundamental moral obligation that applies to all individuals, including citizens. Conversely, the duty to uphold the law requires following legal boundaries, which includes not committing crimes like theft. Finally, the duty to promote justice may lead some to argue that taking action against corruption, even if illegal, is necessary to protect the public interest and prevent further harm.

**</summary>**