PLURALISTIC OFF-POLICY EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized preference alignment for LLMs with diverse human preferences requires evaluation and alignment methods that capture pluralism. Most existing preference alignment datasets are logged under policies that differ substantially from the evaluated LLMs, while existing off-policy estimators focus primarily on overall utility and ignore preference diversity. Extending off-policy evaluation to pluralistic preference alignment is non-trivial, as diversity objectives depend on the target policy distribution and cannot be reliably estimated from logged data using standard Inverse Propensity Scoring (IPS) alone. We propose a framework for offline pluralistic preference evaluation and alignment in LLMs. Our method introduces a unified reward that combines collaborative utility from human preference signals and an entropy-inspired diversity component. To enable practical estimation under logged interactions, we derive decomposable IPS estimators that separately evaluate utility and diversity, and introduce a tractable response-level surrogate that avoids directly optimizing unstable full-slate importance weights. Based on the resulting off-policy value function, our method further enables direct off-policy optimization for improved pluralistic alignment. Empirical results show that our method effectively enhances pluralistic response generation while maintaining competitive downstream generation quality in many settings

1 INTRODUCTION

Preference-aligned language models can generate responses aligned with general human preferences, but recent work shows that they often fail to capture diverse preferences among multiple acceptable responses, motivating pluralistic alignment Sorensen et al. (2024); Chen et al. (2024a); Feng et al. (2024). Ideally, a pluralistically aligned model should prioritize helpful responses while covering diverse, potentially non-majority preferences, which is useful for serving conflicting user needs Jang et al. (2023a) and representing varied perspectives for the same input He-Yueya et al. (2024).

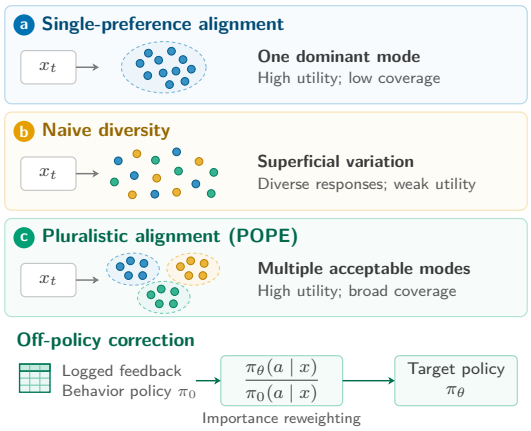


Figure 1: Illustration of pluralistic alignment. (a) Majority preference causes collapse to a single response Rafailov et al. (2023); (b) naive diversity yields superficial variation without preference understanding; (c) POPE covers diverse human-acceptable preference modes with off-policy correction.

A central challenge is to evaluate and optimize two competing desiderata: *collaborative utility*, the aggregate value of responses judged by the user population Zheng et al. (2015), and *diversity*, the range of valid responses the model can cover, analogous to utility–long-tail tradeoffs in fair machine learning and recommender systems Fu et al. (2023); Caton & Haas (2024). Existing pluralistic alignment methods often emphasize matching preference distributions Santurkar et al. (2023); DURMUS et al. (2024); Chen et al. (2024b); Zhao et al. (2024), but may overlook collaborative utility. Conversely, utility-based methods, including reward-model-based alignment Ji et al. (2025) and DPO Rafailov et al. (2023), can bias models toward a narrow set of high-reward responses Sorensen et al. (2024); Fu et al. (2025);

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

Chen et al. (2024c), as also shown in our controlled study (Table 1). Thus, aligning models to diverse preferences while preserving global utility remains an open problem.

This challenge is compounded by the fact that preference datasets are usually logged under behavior policies different from the target model being evaluated or optimized. Although off-policy evaluation has been studied for language model alignment Wu et al. (2025); Bhargava et al. (2024), existing methods mainly estimate global utility and do not explicitly address pluralistic preference evaluation.

We propose **Pluralistic Off-Policy Evaluation** (POPE), a framework for evaluating and optimizing LLMs under both response *utility* and *diversity*. POPE defines a unified reward combining collaborative utility from human feedback, such as upvotes or relevance scores, with an entropy-inspired diversity reward related to response-space coverage. We derive decomposable IPS-based estimators for the utility and diversity components, including a soft-entropy-based diversity estimator inspired by Soft Q-Learning Haarnoja et al. (2017). The resulting decomposable surrogate avoids directly optimizing the high-dimensional set-level IPS objective and enables tractable off-policy optimization of LLM policies via gradient descent. Empirically, POPE improves pluralistic response generation while preserving downstream generation quality.

Our main contributions are as follows:

- We introduce POPE, a framework for pluralistic off-policy evaluation and alignment in LLMs.
- We formulate a unified utility–diversity value function and derive decomposable IPS-based estimators for logged human-feedback data.
- We propose a differentiable off-policy optimization scheme and demonstrate improved pluralistic response generation on in-domain and cross-domain review-generation benchmarks.

2 Problem Formulation

We first formalize pluralistic preference alignment (Section 2), including the policy, reward, and value function, and then present a controlled study (Section 2.1) showing that standard alignment methods such as SFT and DPO struggle to balance collaborative utility and response diversity.

We study a policy that answers a query $x_t \in \mathcal{X}$ using a finite response candidate set $\mathcal{C}_t$ and an ordered slate $S_t = (a_{t,1}, \ldots, a_{t,K})$. For the guarantees below, $\mathcal{C}_t$ is fixed independently of the logged slate and shared by the target and logging policies; it is not the collection of responses revealed only after logging. Each slate position is an independent draw *with replacement*, so repetitions are possible even if $K \le |\mathcal{C}_t|$. This idealized setting formalizes one aspect of pluralistic alignment Sorensen et al. (2024); Jang et al. (2023a); human-written review pools need not follow this sampling model.

We distinguish scalar feedback, such as normalized helpfulness scores, from within-slate rankings. A pairwise ranking can be modeled using Bradley–Terry–Luce preferences Bradley & Terry (1952), and full-slate rankings using Plackett–Luce Plackett (1975); Debreu (1960); neither model by itself makes a rank-derived score an action-local utility. The response-level guarantee below applies to scalar feedback whose conditional mean depends only on the query and response.

Policy. Because open-ended responses vary in length, we use a length-normalized response score. Given an LLM with token-level distribution p_θ, define

$$s_\theta(a \mid x) = \exp\Big(\frac{1}{|a|}\sum_{h=1}^{|a|} \log p_\theta(a_h \mid a_{<h}, x)\Big), \qquad \pi_\theta(a \mid x_t) = \frac{s_\theta(a \mid x_t)}{\sum_{a' \in \mathcal{C}_t} s_\theta(a' \mid x_t)}. \tag{1}$$

For the prespecified finite candidate set $\mathcal{C}_t$, we use $\pi_\theta(a \mid x_t)$ for this finite-candidate distribution. The length-normalized score s_θ is not a normalized probability over unrestricted text, and normalizing over a candidate pool selected from the observed log does not recover generation propensities.

Under the ordered, independent-with-replacement sampling model, the slate probability and full-slate importance ratio are

$$\pi_\theta(S_t \mid x_t) = \prod_{i=1}^{K} \pi_\theta(a_{t,i} \mid x_t), \qquad w_t(S_t) = \frac{\pi_\theta(S_t \mid x_t)}{\pi_0(S_t \mid x_t)} = \prod_{i=1}^{K} \frac{\pi_\theta(a_{t,i} \mid x_t)}{\pi_0(a_{t,i} \mid x_t)}. \tag{2}$$

For this model, the importance ratio is valid only when the true logging probabilities are known. Without-replacement or slate-dependent sampling requires the corresponding joint or marginal inclusion probabilities; merely multiplying single-response probabilities is generally incorrect.

Reward. Following pluralistic alignment Sorensen et al. (2024), we design a two-component reward that encourages both high aggregate human utility and coverage over multiple plausible responses.

Collaborative utility reward. For each response $a_{t,i}$, let $\eta_{t,i} \in [0, 1]$ be its normalized feedback. In the exact estimator result, scalar feedback has an invariant conditional mean $\mathbb{E}[\eta_{t,i} \mid x_t, S_t] = u(x_t, a_{t,i})$, independent of the other slate members. Upvotes or helpfulness scores can serve as empirical signals, although this invariance need not hold for observational feedback. For rank-only feedback we retain the descriptive conversion

$$\eta_{t,i} = \frac{K - r_{t,i}}{K - 1}, \qquad R_{\mathrm{cu}}(x_t, S_t) = \sum_{i=1}^{K} \eta_{t,i}(x_t, a_{t,i}), \qquad K > 1. \tag{3}$$

Here $r_{t,i} \in \{1, \ldots, K\}$ is the rank position of response $a_{t,i}$, with $r_{t,i} = 1$ denoting the most preferred response. The top-ranked response receives 1 and the bottom-ranked response 0, but these scores are *not* action-local: for every complete ranking their sum is $K/2$, regardless of the slate. Consequently, applying the response-level IPS to this heuristic would not give an unbiased estimator of slate utility.

With slate-independent scalar feedback, maximizing expected R_{cu} rewards responses with higher population utility; the rank-only conversion does not provide this guarantee.

Diversity reward. To encourage pluralistic response generation, we use an entropy-inspired diversity reward defined as the aggregate self-information of responses in the slate:

$$R_{\mathrm{div}}(x_t, S_t) = \sum_{i=1}^{K} -\log \pi(a_{t,i} \mid x_t). \tag{4}$$

Its target-policy expectation is the entropy of the finite-candidate *response distribution*. This discourages concentration on a few responses but need not increase coverage of distinct, acceptable preference modes.

Unified pluralistic value. Given n queries and a target policy π, we define

$$V_\lambda(\pi) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{S_t \sim \pi(\cdot|x_t)} \Big[R_{\mathrm{cu}}(x_t, S_t) + \lambda R_{\mathrm{div}}(x_t, S_t) \Big], \tag{5}$$

Here the expectation also includes feedback randomness conditional on the slate when $\eta_{t,i}$ is stochastic; $\lambda \geq 0$ controls the utility–response-entropy trade-off. At $\lambda = 0$ the objective is utility-only; increasing λ need not improve coverage of valid viewpoints.

In our experiments, we set $\lambda = 1$ unless otherwise specified. Because collecting fine-grained online feedback for open-ended responses is costly Gao et al. (2021); Bhargava et al. (2024), we focus on offline evaluation and optimization from logged pluralistic human-response data. Our goal is to estimate and optimize $V_\lambda(\pi)$ using feedback on responses generated by a behavior policy π_0.

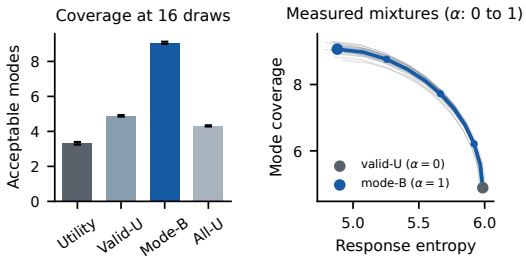


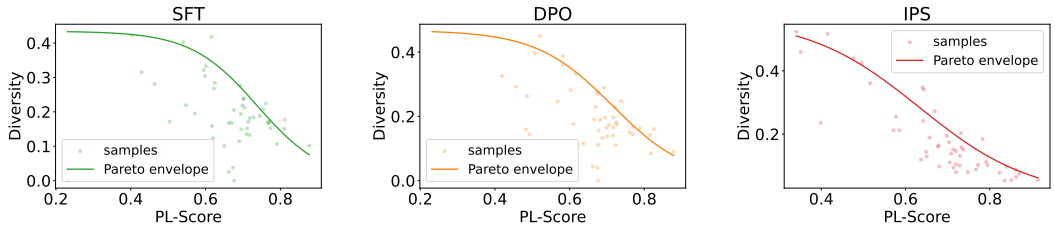
Figure 2: Synthetic experiment of coverage over 20 seeds, 48 contexts, 512 candidates per seed, with $K = 16$. Left: coverage of four policies with 95% seed-bootstrap intervals. Right: evaluated mixtures $\pi_\alpha = (1 - \alpha)\pi_{\mathrm{valid}} + \alpha\pi_{\mathrm{mode}}$. Thin paths show individual seed averages; the bold path shows their mean.

To measure whether a policy represents multiple valid preferences, we group acceptable responses into preference modes $m \in \mathcal{M}$. Let $q_m(x) = \sum_{a \in m} \pi(a \mid x)$ denote the probability mass assigned to mode m. For K independent samples from π, the expected number of distinct acceptable modes represented is $C_K(\pi \mid x) = \sum_{m \in \mathcal{M}} \big[1 - (1 - q_m(x))^K\big]$. Unlike response entropy, this measure directly captures coverage of distinct acceptable modes rather than dispersion over individual responses.

In Figure 2, mode balancing increases acceptable-mode coverage despite reducing response entropy. Its coverage advantage is 4.174 modes on average and is positive in all 20 seeds (range 3.946–4.443).

Table 1: Comparison of Base, SFT, DPO, and POPE on the Alpaca-GPT4 toy dataset using pluralistic evaluation metrics. We report PL-Score, Pluralistic Coverage, and Distributional Alignment; best values are in **bold**, second-best are underlined, and parentheses show the difference relative to the strongest baseline. The symbol ↑ indicates that higher values are better. The improvement achieved by POPE is significant ($p < 0.05$).

Method	PL-Score↑	Pluralistic Coverage↑	Distributional Alignment↑
Base	0.6593	0.5034	0.5692
SFT	0.6579	0.4490	0.5353
DPO	0.6533	0.4694	0.5450
POPE	**0.6842 (+3.8%)**	**0.5510 (+9.5%)**	**0.5757 (+1.1%)**



(a) SFT (b) DPO (c) POPE

Figure 3: Comparison of Pareto fronts under different optimization methods.

2.1 A Controlled Study of Utility–Diversity Trade-offs

Before introducing our framework, we conduct a controlled diagnostic study to illustrate the utility–diversity trade-off faced by standard alignment methods. Rather than fully representing organic human pluralism, this study provides a controlled setting where multiple plausible responses are available for each prompt and their coverage can be quantitatively measured.

Setup. We construct a synthetic evaluation suite from the Alpaca-GPT4 prompt pool Peng et al. (2023b), pairing each prompt with multiple GPT-4o-generated reference responses Hurst et al. (2024) produced under different style or perspective prompts. This yields a controlled approximation of pluralistic responses; details are provided in Section F. We compare four methods: the unmodified base model (**Base**), supervised fine-tuning (**SFT**), direct preference optimization (**DPO**), and our proposed **POPE**. We evaluate them using Plackett–Luce matching score, Pluralistic Coverage, and Distributional Alignment score, with metric definitions in Section D.

Results and Analysis. Table 1 reports the aggregated results under the same generation budget, with K samples drawn per prompt and metrics averaged across prompts. Improvements in parentheses denote gains over the strongest baseline, and significance is assessed using paired tests across prompts with $p < 0.05$. Across all three metrics, POPE outperforms the baselines, suggesting that it better matches the induced Plackett–Luce preference distribution while covering a broader subset of reference responses. In contrast, SFT tends to concentrate on a single high-preference response, while DPO relies on pairwise preference contrasts and may not capture the full spectrum of plausible responses. The supplementary Pareto envelopes in Figure 3 further illustrate the utility–diversity trade-off through PL-Score and sample-level diversity. SFT and DPO exhibit a trade-off where higher PL-Score often corresponds to lower diversity, whereas POPE's envelope shifts toward the upper-right region, indicating a better balance between preference alignment and response diversity. Overall, this controlled study provides motivating evidence that explicitly modeling both utility and diversity can improve pluralistic response generation; since the references are synthetically generated, we further evaluate POPE on real-world logged response datasets in subsequent sections.

3 POPE: PLURALISTIC OFF-POLICY EVALUATION AND ALIGNMENT

Having established the problem formulation and demonstrated empirically that standard methods struggle with the utility–diversity trade-off (Section 2.1), we now present POPE (**P**luralistic **O**ff-**P**olicy **E**valuation and Alignment). POPE is a framework for evaluating and optimizing LLMs from logged human feedback while explicitly balancing collaborative utility and response diversity.

Let $\mathcal{D}_0 = \{(x_t, S_t, \boldsymbol{\eta}_t)\}_{t=1}^n$ denote the idealized logged dataset collected by a known behavior policy π_0, where $S_t = (a_{t,1}, \ldots, a_{t,K})$ is an ordered slate and $\boldsymbol{\eta}_t$ its scalar response feedback. The human-written review datasets use an estimated proxy rather than this known logging policy. Our goal is to estimate and optimize the pluralistic value $V_\lambda(\pi)$ of a target policy π, where $\lambda \geq 0$ controls the trade-off between collaborative utility and diversity. Under the slate-independent scalar-feedback assumption and overlap $\pi_0(a \mid x) > 0$ whenever $\pi(a \mid x) > 0$, the two components of $V_\lambda(\pi)$ admit marginal importance-weighted estimates when the logger is known. The resulting estimator decomposes the pluralistic value into a collaborative-utility component and a diversity component:

$$\hat{V}_{\mathrm{POPE}}(\pi) = \hat{V}_{\mathrm{IPS}}^{\mathrm{cu}}(\pi) + \lambda \hat{V}_{\mathrm{IPS}}^{\mathrm{div}}(\pi). \tag{6}$$

We evaluate both terms on the logged query-slate pairs. Finite-candidate response entropy needs no human annotations and can also be summed exactly over the prespecified candidate set.

3.1 DECOMPOSABLE OFF-POLICY ESTIMATION

A central challenge in pluralistic alignment is that gathering online human feedback for every policy update is prohibitively expensive. We therefore turn to off-policy evaluation (OPE), where the value of a target policy π is estimated using logged interactions collected by a behavior policy π_0. Recall from Section 2 that $V_\lambda(\pi)$ consists of two additive components: collaborative utility and response diversity. This additive structure allows us to estimate the two terms separately before combining them in Equation (6).

Collaborative Utility Estimation. To estimate the collaborative-utility component, we use an IPS estimator over logged responses. For each logged response $a_{t,i}$, define

$$w_{t,i} = \frac{\pi(a_{t,i} \mid x_t)}{\pi_0(a_{t,i} \mid x_t)}, \qquad \hat{V}_{\mathrm{IPS}}^{\mathrm{cu}}(\pi) = \frac{1}{n} \sum_{t=1}^{n} \sum_{i=1}^{K} w_{t,i} \eta_{t,i}(x_t, a_{t,i}). \tag{7}$$

With actual logging propensities and action-local feedback, this estimator corrects for response-sampling shift. Using a language-model proxy for human-written reviews does not inherit this guarantee. In principle, one could also apply IPS at the full-slate level using the slate weight $\pi(S_t \mid x_t)/\pi_0(S_t \mid x_t)$. However, full-slate importance weights can become unstable as the slate size K grows, especially when the target and behavior policies differ substantially. We therefore use the response-level decomposed form in Equation (7), which is consistent with the additive utility reward and the independent response-sampling assumption stated in Section 2. When needed, standard IPS stabilization techniques such as weight clipping or self-normalized IPS can be applied, but we keep the main formulation in its unclipped form for clarity.

Diversity Estimation. For the diversity component, we estimate the entropy-inspired reward from logged responses using importance weighting:

$$\hat{V}_{\mathrm{IPS}}^{\mathrm{div}}(\pi) = \frac{1}{n} \sum_{t=1}^{n} \sum_{i=1}^{K} w_{t,i} \left[-\log \pi(a_{t,i} \mid x_t) \right]. \tag{8}$$

The expectation of $-\log \pi(a \mid x_t)$ under π is its finite-candidate response entropy $H(\pi(\cdot \mid x_t))$. The weighted sum above estimates that expectation from logged responses, without obtaining new human annotations. Because the normalized finite-candidate probabilities are already available, $\sum_{a \in \mathcal{C}_t} -\pi(a \mid x_t) \log \pi(a \mid x_t)$ is also computable exactly.

3.2 Off-Policy Evaluation and Alignment

A direct full-slate IPS estimator would reweight each logged slate by $\pi(S_t \mid x_t)/\pi_0(S_t \mid x_t)$. Under the independent response-sampling assumption introduced in Section 2, this ratio factorizes as

$$\frac{\pi(S_t \mid x_t)}{\pi_0(S_t \mid x_t)} = \prod_{i=1}^{K} \frac{\pi(a_{t,i} \mid x_t)}{\pi_0(a_{t,i} \mid x_t)}. \tag{9}$$

The additive rewards instead permit an exact marginalization in expectation, avoiding products of importance weights in the empirical training objective.

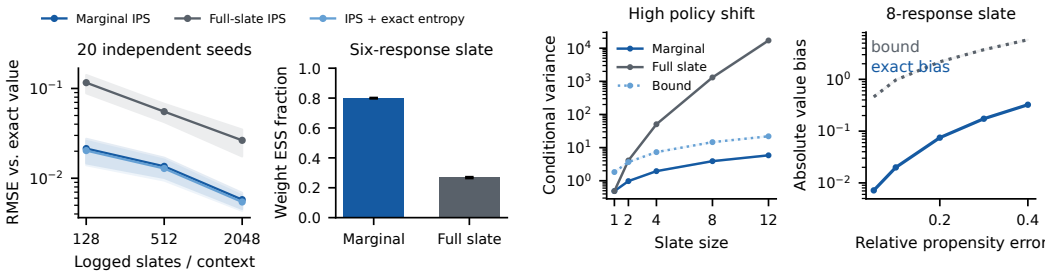


Figure 4: Theory-aligned diagnostics. Left: off-policy estimation error and effective sample size. Right: conditional variance and proxy-propensity sensitivity.

Proposition 1 (Exact marginal POPE identity)**.** *Fix the queries and a target policy π independently of the evaluation log. Suppose each logged slate contains K independent draws from the true, overlapping logger π_0 on a prespecified finite $\mathcal{C}_t$, and $\mathbb{E}[\eta_{t,i} \mid x_t, S_t] = u(x_t, a_{t,i}) \in [0,1]$, independently of other slate members. Then*

$$\widehat{J}_{\mathrm{POPE}}(\pi) \quad = \frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{K} \frac{\pi(a_{t,i} \mid x_t)}{\pi_0(a_{t,i} \mid x_t)} \cdot \left[\eta_{t,i}(x_t, a_{t,i}) - \lambda \log \pi(a_{t,i} \mid x_t)\right], \tag{10}$$

where $\widehat{J}_{\mathrm{POPE}}(\pi) = \hat{V}_{\mathrm{POPE}}(\pi)$, and $\mathbb{E}_{\mathcal{D}_0}[\widehat{J}_{\mathrm{POPE}}(\pi) \mid x_{1:n}] = V_\lambda(\pi)$.

In Figure 4, all three estimators have the same exact expectation for every one of the 960 fixed-policy contexts; finite-sample error falls as log size grows. At six draws, population weight effective-sample-size fractions average 0.800 for marginal weights versus 0.268 for slate weights. This is not a pointwise lower bound or a guarantee for rank-only feedback: a separate exact pairwise example yields true rank-sum value 1 but expected marginal IPS 1.3 (Section G).

Proposition 2 (Stability and propensity sensitivity)**.** *Under 1, assume action–feedback pairs are also independent across slate positions, $|\eta_i - \lambda \log \pi(a_i \mid x)| \le R$, and put $c = \sum_{a \in \mathcal{C}(x)} \pi(a \mid x)^2/\pi_0(a \mid x)$. For one query, write $\hat{v}_{\mathrm{marg}} = \sum_i w_i[\eta_i - \lambda \log \pi(a_i \mid x)]$ and $W = \prod_i w_i$. Then*

$$\mathrm{Var}(\hat{v}_{\mathrm{marg}} \mid x) \le KR^2c, \qquad \mathbb{E}_{\pi_0}[W^2 \mid x] = c^K. \tag{11}$$

If a fixed or independently estimated proxy $\hat{\pi}_0$ satisfies $|\hat{\pi}_0(a \mid x)/\pi_0(a \mid x) - 1| \le \epsilon < 1$ for all target-supported candidates, its plug-in marginal value obeys

$$|\mathbb{E}_{\pi_0}[\hat{v}_{\mathrm{proxy}} \mid x] - V_{\lambda,x}(\pi)| \le \frac{KR\epsilon}{1-\epsilon}. \tag{12}$$

In Figure 4, we further examine the stability of the estimator. The second moment of the slate weight grows as c^K, although this fact alone does not imply a universal ordering of the value-estimator variances. At a relative propensity error of 0.2, the mean absolute value bias is 0.075, remaining below the mean bound of 2.180. Removing the 5% of candidates with the largest target-value contributions leaves, on average, 1.574 of the $K = 8$ target value unidentifiable; for the human-review proxy, no corresponding ϵ guarantee is available.

3.3 Pluralistic Alignment via Policy Gradient

The decomposable POPE surrogate in Equation (10) is differentiable with respect to the policy parameters θ. We therefore align the LLM with pluralistic preferences by maximizing $\widehat{J}_{\mathrm{POPE}}(\pi_\theta)$ via gradient ascent:

$$\theta \leftarrow \theta + \alpha \nabla_\theta \widehat{J}_{\mathrm{POPE}}(\pi_\theta). \tag{13}$$

The gradient takes the following form:

$$\begin{aligned} \nabla_\theta \widehat{J}_{\mathrm{POPE}}(\pi_\theta) = \frac{1}{n} \sum_{t=1}^{n} \sum_{i=1}^{K} \frac{\pi_\theta(a_{t,i} \mid x_t)}{\pi_0(a_{t,i} \mid x_t)} \nabla_\theta \log \pi_\theta(a_{t,i} \mid x_t) \\ \cdot \left[\eta_{t,i}(x_t, a_{t,i}) - \lambda \left(\log \pi_\theta(a_{t,i} \mid x_t) + 1 \right) \right]. \end{aligned} \tag{14}$$

In this update rule, $\eta_{t,i}$ provides the collaborative-utility signal from logged human feedback, while the term $-\lambda(\log \pi_\theta(a_{t,i} \mid x_t) + 1)$ acts as an adaptive entropy-inspired regularizer. This update increases weight on high-feedback responses while promoting response entropy; it does not by itself guarantee coverage of valid minority modes. The derivative of $\log \pi_\theta$ includes the normalizer over $\mathcal{C}_t$ in Equation (1), while the logging propensity is fixed. The identity in 1 holds for a fixed target policy; maximizing its noisy estimate on the same logs does not make the selected policy's in-sample score an unbiased evaluation.

4 Experiments

We evaluate POPE on controlled and real-world review-generation benchmarks, covering in-domain movie reviews, cross-domain transfer, pluralistic coverage, and auxiliary generation-quality analyses. We first evaluate POPE's impact on off-policy value $\hat{V}_{\mathrm{POPE}}(\theta)$, response quality, and diversity on in-domain movie-review datasets (Section 4). We then test cross-domain transfer on Amazon Music and Amazon Video Games (Section 4), examine pluralistic coverage across datasets (Section 4), and analyze additional generation-quality metrics using AlignScore and Self-BLEU (Section I).

Datasets. We evaluate POPE on a controlled Alpaca-GPT4-based benchmark Peng et al. (2023a) (Section 2.1) and real-world review datasets. Reddit Movie Reviews He et al. (2023) and Amazon Movies Reviews Hou et al. (2024) are used for in-domain evaluation (Section 4), while Amazon Music and Video Games Reviews Hou et al. (2024) test cross-domain transfer (Section 4). In the review datasets, item descriptions serve as queries and user reviews with helpfulness votes provide logged responses and preference signals. Additional pluralistic-coverage and generation-quality evaluations are reported in Sections 4 and I.

Models and Baselines. We evaluate Llama 3 1B/3B Dubey et al. (2024), Phi-3.5-mini 3.8B Abdin et al. (2024), and Qwen 3 4B Team (2025). We compare **Base**, **SFT**, **DPO** Rafailov et al. (2023), and **POPE**, where the latter optimizes our off-policy objective over logged candidate responses.

Training Setup. Because human-written reviews do not provide true logging propensities, we estimate the behavior policy using a fixed GPT2-Medium proxy Radford et al. (2019) and compute importance weights over each finite candidate set. For review experiments, POPE, SFT, and DPO are trained on the Reddit validation split and evaluated both in-domain and cross-domain. Full data construction, propensity estimation, metrics, and implementation details are provided in the appendix.

In-domain Movie Review Experiments. We evaluate whether POPE improves pluralistic alignment on two in-domain movie-review datasets: Amazon Movies and Reddit. Table 2 reports helpfulness (Hel.), relevance (Rel.), unigram and bigram lexical diversity (D-1, D-2), and the estimated off-policy value $\hat{V}_{\mathrm{POPE}}(\theta)$ for **Base**, **SFT**, **DPO**, and **POPE**. Overall, POPE consistently improves the estimated pluralistic off-policy value and often improves lexical diversity while maintaining competitive helpfulness and relevance. In contrast, SFT tends to reduce diversity and sometimes degrade relevance, suggesting over-concentration on narrow high-vote responses, while DPO preserves relevance in several cases but yields limited gains in diversity and pluralistic value. These results suggest that combining collaborative utility with entropy-inspired diversity improves in-domain pluralistic response generation.

Cross-domain Transfer Experiments. To evaluate cross-domain transfer within review-generation, we apply models trained on source review data without additional adaptation to Amazon Music and

Table 2: Comparison results of Base LLMs (Base), supervised fine-tuning (SFT), direct preference optimization (DPO), and POPE on two in-domain review datasets (Amazon Movies and Reddit). We report Helpfulness (Hel.), Relevance (Rel.), Distinct-1 (D-1), Distinct-2 (D-2), and off-policy evaluation $\hat{V}(\theta)$. Best results are in **bold**, second-best underlined. ↑ indicates higher is better. The improvement achieved by POPE is significant ($p < 0.05$).

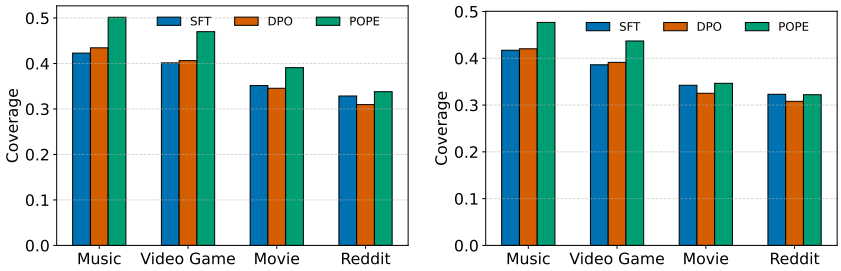
Model	Method	Amazon Movies					Reddit				
		Hel.↑	Rel.↑	D-1↑	D-2↑	$\hat{V}(\theta)$↑	Hel.↑	Rel.↑	D-1↑	D-2↑	$\hat{V}(\theta)$↑
Llama3	Base	**29.17**	**57.71**	14.35	43.16	0.3569	23.77	55.57	**10.66**	35.43	0.4171
	SFT	21.79	38.65	11.55	31.12	0.3525	19.24	33.29	7.08	22.45	0.3952
	DPO	28.85	57.57	**14.66**	**44.04**	0.3567	**23.99**	55.92	10.51	35.11	0.4852
	POPE	25.54	48.38	14.08	43.85	**0.4008**	23.19	**58.32**	10.44	**36.55**	**0.8513**
Qwen3	Base	26.71	49.44	10.68	32.47	0.3528	25.42	63.04	7.15	28.52	0.4218
	SFT	27.47	40.70	8.48	22.53	0.3382	22.97	44.56	4.36	14.53	0.4029
	DPO	26.93	**60.78**	**15.49**	**46.42**	0.3526	25.42	63.36	7.18	28.56	0.4961
	POPE	**28.74**	50.21	12.10	39.05	**0.3778**	25.10	**69.47**	**8.20**	**34.22**	**0.5916**
Phi-3.5	Base	25.89	59.04	15.31	45.47	0.4682	23.10	60.36	10.89	37.96	0.5847
	SFT	25.10	43.79	8.29	21.89	0.4616	20.34	50.68	4.85	15.76	0.4092
	DPO	26.99	48.99	10.78	32.63	0.4680	23.26	**60.37**	11.01	38.38	0.4722
	POPE	**29.52**	**67.49**	**18.04**	**55.23**	**0.5631**	**24.83**	60.26	**11.58**	**40.74**	**0.8667**

Table 3: Comparison of Base LLMs (Base), supervised fine-tuning (SFT), direct preference optimization (DPO), and POPE on two cross-domain review datasets (Amazon Music & Amazon Video Game). We report Helpfulness (Hel.), Relevance (Rel.), Distinct-1 (D-1), Distinct-2 (D-2) and policy evaluation $\hat{V}(\theta)$; best in **bold**, second-best underlined. The symbol ↑ indicates that higher values are better. The improvement achieved by POPE is significant ($p < 0.05$).

Model	Method	Amazon Music					Amazon Video Game				
		Hel.↑	Rel.↑	D-1↑	D-2↑	$\hat{V}(\theta)$↑	Hel.↑	Rel.↑	D-1↑	D-2↑	$\hat{V}(\theta)$↑
Llama3	Base	36.86	66.28	8.14	32.37	0.4144	**34.92**	**65.76**	6.88	29.77	0.3945
	SFT	25.67	43.37	6.77	22.81	0.4089	23.12	41.22	6.27	22.89	0.3880
	DPO	**36.90**	**66.45**	8.17	32.56	0.4141	34.59	65.46	6.87	29.77	0.3943
	POPE	33.78	60.81	**8.30**	**34.27**	**0.4321**	32.45	59.61	**7.09**	**31.22**	**0.4219**
Qwen3	Base	32.49	58.87	6.87	26.81	**0.4248**	29.58	55.59	5.65	24.06	0.4029
	SFT	32.95	55.97	5.43	18.06	0.4024	8.62	51.19	4.43	15.96	0.3817
	DPO	32.82	59.38	6.79	26.56	0.4246	29.74	55.97	5.67	24.02	0.4027
	POPE	**34.21**	**60.89**	**7.47**	**31.97**	0.4175	**31.57**	**58.39**	**6.29**	**28.43**	**0.4117**
Phi-3.5	Base	38.92	73.16	9.34	39.81	0.4311	35.40	70.70	8.13	36.01	0.4189
	SFT	37.84	58.93	4.90	17.69	0.4668	35.03	60.11	4.83	18.57	0.4664
	DPO	38.91	73.40	9.31	39.44	0.4680	35.57	70.80	8.16	35.97	0.4707
	POPE	**41.51**	**78.57**	**11.17**	**46.89**	**0.5398**	**38.27**	**76.91**	**9.78**	**43.33**	**0.5657**

Amazon Video Games Ni et al. (2019), which differ from the movie-review source domain but remain within the same task family. Table 3 reports helpfulness (Hel.), relevance (Rel.), lexical diversity (D-1, D-2), and estimated pluralistic off-policy value $\hat{V}_{\mathrm{POPE}}(\theta)$. Overall, POPE generally improves the utility–diversity trade-off compared with SFT and DPO, with the strongest gains on Phi-3.5, where it improves response quality, diversity, and estimated pluralistic value on both target domains. Results for Llama3 and Qwen3 are more mixed: POPE often improves diversity and $\hat{V}_{\mathrm{POPE}}(\theta)$, but does not uniformly dominate all baselines on helpfulness or relevance. These results suggest that POPE can improve pluralistic coverage under domain shift, while its effectiveness depends on the backbone model and target review domain. Since all datasets remain within review generation, we interpret the results as evidence of cross-domain transfer within this task family rather than generalization to structurally different open-ended generation tasks. Implementation details are provided in Section E.

Pluralistic Evaluation Across Datasets. We measure the Pluralistic Coverage of generated responses on four review datasets (Amazon Music, Amazon Video Games, Amazon Movies, and Reddit datasets), using both the Llama-3 and Phi-3.5 model families. Pluralistic Coverage quantifies the fraction of distinct human opinion categories that the model outputs replicate (details in Section D). We compare three alignment methods: SFT, DPO, and our proposed POPE. As shown in Figures 5a and 5b, POPE consistently achieves the highest coverage on every dataset. On Llama-3, POPE improves coverage by around 15–20% relative to SFT and by 10–15% relative to DPO. On Phi-3.5,



(a) Llama3 Coverage. (b) Phi3.5 Coverage.

Figure 5: Pluralistic evaluation comparison across methods (SFT, DPO, POPE) on four datasets: Music, Video Game, Movie, and Reddit.

POPE delivers gains of approximately 10–15% over both baselines. These results demonstrate that our off-policy pluralistic alignment effectively enables LLMs to capture a broader range of human preferences across domains. POPE thus provides a robust, domain-agnostic mechanism for enhancing pluralism in open-ended response generation.

5 Related Work

Pluralistic Alignment. Standard alignment methods learn from aggregate human preferences Rafailov et al. (2023); Ouyang et al. (2022); Stiennon et al. (2022), despite substantial preference heterogeneity Jang et al. (2023b); Zhang et al. (2024a). Pluralistic alignment instead seeks to represent diverse preferences Sorensen et al. (2024); Xie et al. (2025), either through user- or group-conditioned models Lee et al. (2024); Wang et al. (2024); Kumar et al. (2024); Poddar et al. (2024); Balepur et al. (2025) or distributional preference objectives Yao et al. (2025); Li et al. (2024a); Siththaranjan et al. (2024). These approaches generally assume explicit conditioning information or focus on preference modeling and interactive optimization. In contrast, POPE studies offline pluralistic alignment from logged responses, jointly estimating collaborative utility and response diversity.

Offline Policy Evaluation and Optimization. Recent work adapts off-policy evaluation to LLM alignment using importance sampling, doubly robust estimation, and model-based estimators over logged feedback Bhargava et al. (2024); Gao et al. (2023); Li et al. (2024b), while OCEAN applies inverse propensity weighting to offline reasoning traces Wu et al. (2024). Related offline preference-optimization methods use logged data to construct or reweight training targets without new human queries Ji et al. (2024); Zhou et al. (2024); Yin et al. (2024). These methods primarily target aggregate utility or task performance, whereas POPE explicitly decomposes offline pluralistic alignment into utility and diversity objectives.

Diversity in Recommendation. Recommender systems similarly move beyond accuracy by considering diversity, serendipity, and fairness Duricic et al. (2023); Knees et al. (2023); Jafari Ziarani & Ravanmehr (2021); Zhao et al. (2023b). Prior work studies personalized diversity–serendipity trade-offs Yin & Zhao (2024) and mechanisms for pluralistic exposure while preserving user control Vermeulen (2022). These objectives motivate our treatment of diversity as a first-class component alongside collaborative utility.

6 Conclusion

In this work, we present Pluralistic Off-Policy Evaluation (POPE), an offline framework for pluralistic evaluation and alignment from logged human-feedback data. POPE combines collaborative utility, capturing aggregate preference signals, with an entropy-inspired diversity term encouraging coverage of plausible responses. We derive decomposable IPS-style estimators for both components and use the resulting response-level surrogate for proxy off-policy optimization. Across controlled and review-generation benchmarks, POPE consistently improves estimated pluralistic value and diversity metrics while maintaining competitive helpfulness, relevance, and generation quality. These results highlight the importance of explicitly modeling the utility–diversity trade-off and motivate future work on propensity estimation, human-centered evaluation, and broader open-ended generation tasks.

AI USE STATEMENT

We use AI tools for figure preparation and writing edits.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. *URL https://arxiv. org/abs/2404.14219*, 2024.

Nishant Balepur, Vishakh Padmakumar, Fumeng Yang, Shi Feng, Rachel Rudinger, and Jordan Lee Boyd-Graber. Whose boat does it float? improving personalization in preference tuning via inferred user personas. *arXiv preprint arXiv:2501.11549*, 2025.

Aniruddha Bhargava, Lalit Jain, Branislav Kveton, Ge Liu, and Subhojyoti Mukherjee. Off-policy evaluation from logged human feedback. *arXiv preprint arXiv:2406.10030*, 2024.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7), April 2024. ISSN 0360-0300. doi: 10.1145/3616865. URL https://doi.org/10.1145/3616865.

Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Korlakai Vinayak. PAL: pluralistic alignment framework for learning from heterogeneous preferences. *CoRR*, abs/2406.08469, 2024a.

Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Korlakai Vinayak. Pal: Pluralistic alignment framework for learning from heterogeneous preferences, 2024b. URL https://arxiv.org/abs/2406.08469.

Lichang Chen, Chen Zhu, Davit Soselia, Jiuhai Chen, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. Odin: Disentangled reward mitigates hacking in rlhf. *arXiv preprint arXiv:2402.07319*, 2024c.

Gerard Debreu. Individual choice behavior: A theoretical analysis, 1960.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.

Tomislav Duricic, Dominik Kowald, Emanuel Lacic, and Elisabeth Lex. Beyond-accuracy: A review on diversity, serendipity and fairness in recommender systems based on graph neural networks. *arXiv preprint arXiv:2310.02294*, 2023. URL https://arxiv.org/abs/2310.02294.

Esin DURMUS, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=zl16jLb91v.

Weike Fang and Meng Jiang. Investigating relationships between accuracy and diversity in multireference text generation. 2022.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. Modular pluralism: Pluralistic alignment via multi-llm collaboration. *arXiv preprint arXiv:2406.15951*, 2024.

Jiayi Fu, Xuandong Zhao, Chengyuan Yao, Heng Wang, Qi Han, and Yanghua Xiao. Reward shaping to mitigate reward hacking in rlhf. *arXiv preprint arXiv:2502.18770*, 2025.

Zhe Fu, Xi Niu, and Mary Lou Maher. Deep learning models for serendipity recommendations: A survey and new perspectives. *ACM Comput. Surv.*, 56(1), August 2023. ISSN 0360-0300. doi: 10.1145/3605145. URL https://doi.org/10.1145/3605145.

Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten De Rijke, and Tat-Seng Chua. Advances and challenges in conversational recommender systems: A survey. *AI open*, 2:100–126, 2021.

Qitong Gao, Ge Gao, Juncheng Dong, Vahid Tarokh, Min Chi, and Miroslav Pajic. Off-policy evaluation for human feedback. *Advances in Neural Information Processing Systems*, 36:9065–9091, 2023.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pp. 1352–1361. PMLR, 2017.

Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian J. McAuley. Large language models as zero-shot conversational recommenders. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, pp. 720–730. ACM, 2023.

Joy He-Yueya, Wanjing Anya Ma, Kanishk Gandhi, Benjamin W. Domingue, Emma Brunskill, and Noah D. Goodman. Psychometric alignment: Capturing human knowledge distributions via language models, 2024. URL https://arxiv.org/abs/2407.15645.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian J. McAuley. Bridging language and items for retrieval and recommendation. *CoRR*, abs/2403.03952, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Sagar Imambi, Kolla Bhanu Prakash, and GR Kanagachidambaresan. Pytorch. *Programming with TensorFlow: solution for edge computing applications*, pp. 87–104, 2021.

Reza Jafari Ziarani and Reza Ravanmehr. Serendipity in recommender systems: A systematic literature review. *Journal of Computer Science and Technology*, 36(1):119–138, 2021. doi: 10.1007/s11390-020-0135-9. URL https://link.springer.com/article/10.1007/s11390-020-0135-9.

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

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

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.

Xiang Ji, Sanjeev Kulkarni, Mengdi Wang, and Tengyang Xie. Self-play with adversarial critic: Provable and scalable offline alignment for language models. *arXiv preprint arXiv:2406.04274*, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Peter Knees, Julia Neidhardt, and Irina Nalis. Recommender systems: Techniques, effects, and measures toward pluralism and fairness. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor (eds.), *Recommender Systems Handbook*. Springer, 2023. doi: 10.1007/978-3-031-45304-5_27. URL https://link.springer.com/chapter/10.1007/978-3-031-45304-5_27.

Sachin Kumar, Chan Young Park, Yulia Tsvetkov, Noah A Smith, and Hannaneh Hajishirzi. Compo: Community preferences for language model personalization. *arXiv preprint arXiv:2410.16027*, 2024.

Seongyun Lee, Sue Hyun Park, Seungone Kim, and Minjoon Seo. Aligning to thousands of preferences via system message generalization. *arXiv preprint arXiv:2405.17977*, 2024.

Dexun Li, Cong Zhang, Kuicai Dong, Derrick Goh Xin Deik, Ruiming Tang, and Yong Liu. Aligning crowd feedback via distributional preference reward modeling, 2024a. URL https://arxiv.org/abs/2402.09764.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

Zihao Li, Xiang Ji, Minshuo Chen, and Mengdi Wang. Policy evaluation for reinforcement learning from human feedback: A sample complexity analysis. In *International Conference on Artificial Intelligence and Statistics*, pp. 2737–2745. PMLR, 2024b.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 188–197, 2019.

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

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with GPT-4. *CoRR*, abs/2304.03277, 2023a.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023b.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. *arXiv preprint arXiv:2408.10075*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=0tWTxYYPnW.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: a roadmap to pluralistic alignment. ICML'24. JMLR.org, 2024.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2022. URL https://arxiv.org/abs/2009.01325.

Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Judith Vermeulen. To nudge or not to nudge: News recommendation as a tool to achieve online media pluralism. *Journal of Media Law*, X(X):1–18, 2022. doi: 10.1080/21670811.2021.2021804. URL https://doi.org/10.1080/21670811.2021.2021804.

Danqing Wang, Kevin Yang, Hanlin Zhu, Xiaomeng Yang, Andrew Cohen, Lei Li, and Yuandong Tian. Learning personalized alignment for evaluating open-ended text generation, 2024. URL https://arxiv.org/abs/2310.03304.

Junda Wu, Xintong Li, Ruoyu Wang, Yu Xia, Yuxin Xiong, Jianing Wang, Tong Yu, Xiang Chen, Branislav Kveton, Lina Yao, et al. Ocean: Offline chain-of-thought evaluation and alignment in large language models. *arXiv preprint arXiv:2410.23703*, 2024.

Junda Wu, Xintong Li, Ruoyu Wang, Yu Xia, Yuxin Xiong, Jianing Wang, Tong Yu, Xiang Chen, Branislav Kveton, Lina Yao, Jingbo Shang, and Julian McAuley. OCEAN: Offline chain-of-thought evaluation and alignment in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=rlgplAuN2p.

Zhouhang Xie, Junda Wu, Yiran Shen, Yu Xia, Xintong Li, Aaron Chang, Ryan Rossi, Sachin Kumar, Bodhisattwa Prasad Majumder, Jingbo Shang, Prithviraj Ammanabrolu, and Julian McAuley. A survey on personalized and pluralistic preference alignment in large language models, 2025. URL https://arxiv.org/abs/2504.07070.

Binwei Yao, Zefan Cai, Yun-Shiuan Chuang, Shanglin Yang, Ming Jiang, Diyi Yang, and Junjie Hu. No preference left behind: Group distributional preference optimization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bgpNJBD6Va.

Kexin Yin and Junqi Zhao. Diversity and serendipity preference-aware recommender system. *Journal of Computer Engineering & Computer Technology*, 2024. doi: 10.47852/bonviewJCCE42023272. URL https://doi.org/10.47852/bonviewJCCE42023272.

Yueqin Yin, Zhendong Wang, Yujia Xie, Weizhu Chen, and Mingyuan Zhou. Self-augmented preference optimization: Off-policy paradigms for language model alignment. *arXiv preprint arXiv:2405.20830*, 2024.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*, 2023.

Michael JQ Zhang, Zhilin Wang, Jena D. Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. Diverging preferences: When do annotators disagree and do models know?, 2024a. URL https://arxiv.org/abs/2410.14632.

Tianhui Zhang, Bei Peng, and Danushka Bollegala. Improving diversity of commonsense generation by large language models via in-context learning. *arXiv preprint arXiv:2404.16807*, 2024b.

Siyan Zhao, John Dang, and Aditya Grover. Group preference optimization: Few-shot alignment of large language models. *arXiv preprint arXiv:2310.11523*, 2023a.

Siyan Zhao, John Dang, and Aditya Grover. Group preference optimization: Few-shot alignment of large language models, 2024. URL https://arxiv.org/abs/2310.11523.

Yuying Zhao, Yu Wang, Yunchao Liu, Xueqi Cheng, Charu C. Aggarwal, and Tyler Derr. Fairness and diversity in recommender systems: A survey. *ACM Transactions on Intelligent Systems and Technology*, 2023b. URL https://arxiv.org/abs/2307.04644. preprint arXiv:2307.04644.

Qianru Zheng, Chi-Kong Chan, and Horace HS Ip. An unexpectedness-augmented utility model for making serendipitous recommendation. In *Advances in Data Mining: Applications and Theoretical Aspects: 15th Industrial Conference, ICDM 2015, Hamburg, Germany, July 11-24, 2015, Proceedings 15*, pp. 216–230. Springer, 2015.

Wenxuan Zhou, Ravi Agrawal, Shujian Zhang, Sathish Reddy Indurthi, Sanqiang Zhao, Kaiqiang Song, Silei Xu, and Chenguang Zhu. Wpo: Enhancing rlhf with weighted preference optimization. *arXiv preprint arXiv:2406.11827*, 2024.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 1097–1100, 2018.

A Review-generation Benchmark Details

Our real-world evaluation focuses on four review-generation datasets: Amazon Movies, Reddit Movie Reviews, Amazon Music, and Amazon Video Games. These datasets provide a practical testbed for pluralistic response generation because each item or query can be associated with multiple user-written reviews and preference signals such as helpfulness votes. This setting allows us to evaluate whether the utility–diversity trade-off studied in the controlled Alpaca-GPT4 setting (Section 2.1) also appears in logged user-review data.

At the same time, all datasets belong to the broader genre of product or user-review generation. Therefore, our experiments should be interpreted as evidence of in-domain and cross-domain transfer within review-generation tasks, rather than as a comprehensive evaluation of all open-ended generation settings such as question answering or dialogue.

B Synthetic Benchmark and Review Dataset Details

For the controlled motivating study (Section 2.1), we construct a subset of the Alpaca-GPT4 instruction-response dataset Peng et al. (2023a), sampling 500 distinct prompts. For each prompt, we use GPT-4o to generate five stylistically distinct reference responses following predefined style instructions: Formal Academic, Casual Conversational, Technical Bullet-Point Overview, Metaphorical/Poetic Description, and Humorous Anecdotal Narrative. These references provide a controlled approximation of pluralistic responses, allowing us to test whether a model can cover multiple plausible response modes under an idealized setting.

This synthetic benchmark is not intended to replace organically collected human preference data. Because the reference responses are generated by a single LLM under predefined style prompts, they may capture stylistic diversity more directly than deeper forms of human preference pluralism. We therefore use this benchmark as a diagnostic study of the utility–diversity trade-off and complement it with evaluations on real-world logged review datasets.

For the real-world review-generation experiments, we use Reddit Movie Reviews He et al. (2023), Amazon Movies Reviews Hou et al. (2024), Amazon Music Reviews, and Amazon Video Games Reviews Hou et al. (2024). These datasets are well suited for studying pluralistic response generation because a single item can receive multiple valid reviews reflecting different user perspectives. At the same time, all four datasets belong to the broader genre of user-review generation. Therefore, our results should be interpreted as evidence of in-domain and cross-domain transfer within this task family rather than as a comprehensive evaluation of all open-ended generation settings.

C Behavior Policy and Propensity Estimation

A practical challenge in applying IPS to review datasets is that the logged responses are human-written reviews rather than samples from an explicitly recorded LLM policy. Thus, the true behavior policy $\pi_0(a_{t,i} \mid x_t)$ is not directly available. Following common practice in off-policy evaluation with implicit logging policies, we approximate π_0 with a fixed proxy language model. Specifically, we use GPT2-Medium Radford et al. (2019) as a behavior-policy proxy and compute the length-normalized likelihood of each logged response conditioned on the item description. For each query x_t, we normalize these scores over the finite candidate response set $\mathcal{C}_t$:

$$\hat{\pi}_0(a_{t,i} \mid x_t) = \frac{s_0(a_{t,i} \mid x_t)}{\sum_{a \in \mathcal{C}_t} s_0(a \mid x_t)}, \qquad s_0(a \mid x_t) = \exp\left(\frac{1}{|a|}\sum_{h=1}^{|a|} \log p_0(a_h \mid a_{<h}, x_t)\right). \quad (15)$$

where p_0 denotes the token distribution of the proxy behavior model. We compute the target-policy propensity $\pi_\theta(a_{t,i} \mid x_t)$ analogously using the corresponding target LLM and normalize it over the same candidate set $\mathcal{C}_t$. The resulting finite-candidate propensities are then used to form the response-level importance weights in POPE.

This proxy-based propensity estimation introduces an approximation: $\hat{\pi}_0$ should be interpreted as a model-based estimate of the implicit human logging distribution, not as the true probability with which a human reviewer wrote a response. We use the same propensity-estimation procedure for all methods and datasets to ensure a consistent comparison.

D Evaluation Metrics Details

Alongside the off-policy evaluation framework presented in Section 3.1, we employ both established and newly proposed, task-specific metrics to comprehensively assess our models' capabilities:

D.1 RQ1 Evaluation Metrics

D.1.1 PL-Score (Higher is better)

To measure how well LLM's response rankings match human Plackett–Luce preference judgments. Let M be the number of ground-truth responses $\{\mathrm{gt}_k\}_{k=1}^{M}$ with up-votes u_k, and N the number of generated candidates $\{\mathrm{gen}_j\}_{j=1}^{N}$. We define,

$$\theta_k = \frac{u_k}{\sum_{\ell=1}^{M} u_\ell}, \quad S_{j,k} = \cos\big(\mathbf{e}(\mathrm{gen}_j),\, \mathbf{e}(\mathrm{gt}_k)\big). \quad (16)$$

Then

$$\text{PL-Score} = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{M} \theta_k\, S_{j,k}. \quad (17)$$

D.1.2 PLURALISTIC COVERAGE (HIGHER IS BETTER)

We define the Pluralistic Coverage as the fraction of distinct reference responses retrieved among the top-k outputs above a similarity threshold δ. Formally,

$$\text{Coverage} = \frac{1}{M}\sum_{k=1}^{M}\mathbb{I}\big(\max_{1\le j\le N} S_{j,k} > \delta\big). \tag{18}$$

In this paper, we adopt δ as 0.8.

D.1.3 DISTRIBUTIONAL ALIGNMENT (HIGHER IS BETTER)

Normalized entropy of the implicit distribution over ground-truth responses, measuring diversity of the model's slate. Let:

$$p_k = \frac{\exp\big(\sum_{j=1}^{N} S_{j,k}/\tau\big)}{\sum_{\ell=1}^{M}\exp\big(\sum_{j=1}^{N} S_{j,\ell}/\tau\big)} \quad (\tau \text{ is a temperature}). \tag{19}$$

Then Distributional Alignment (DA) is defined as:

$$\text{DA} = -\frac{1}{\log M}\sum_{k=1}^{M} p_k \log p_k \tag{20}$$

We set τ as 0.5 in this paper. It is worth noting that distributional alignment quantifies the normalized Shannon entropy of the model's induced preference distribution over human references, akin to how Jensen–Shannon divergence (JSD) Zhao et al. (2023a) measures distributional differences in prior work.

D.1.4 DIVERSITY (HIGHER IS BETTER)

We measure the average pairwise diversity of the generated candidates Fang & Jiang (2022); Zhang et al. (2024b).

$$\text{Diversity} = 1 - \frac{1}{C\binom{N}{2}}\sum_{1\le i<j\le N}\cos\big(\mathbf{e}(\text{gen}_i),\, \mathbf{e}(\text{gen}_j)\big). \tag{21}$$

D.2 RQ2 AND RQ3 EVALUATION METRICS

D.2.1 HELPFULNESS SCORE (HIGHER IS BETTER)

For each model response i, let its human replies be $\{\text{reply}_{i,k}\}_{k=1}^{K_i}$ with raw up-votes $v_{i,k}$. After min–max scaling to $[0, 10]$ to obtain $v'_{i,k}$ and normalizing to weights $w_{i,k} = v'_{i,k}/\sum_j v'_{i,j}$, we define

$$\text{Helpfulness}_i = \sum_{k=1}^{K_i} w_{i,k}\,\cos\big(\mathbf{e}(\text{resp}_i),\, \mathbf{e}(\text{reply}_{i,k})\big). \tag{22}$$

D.2.2 RELEVANCE SCORE (HIGHER IS BETTER)

Cosine similarity between question (or query) and response embeddings Schütze et al. (2008) : $\text{Relevance}_i = \cos(\mathbf{e}(\text{que}_i), \mathbf{e}(\text{resp}_i))$, where embeddings are obtained using SentenceBERT Reimers & Gurevych (2019).

D.2.3 DISTINCT-N (HIGHER IS BETTER)

Distinct-N measures lexical diversity of the generated responses Li et al. (2015).

D.3 RQ4 EVALUATION METRICS

Coverage (higher is better). Coverage measures the fraction of unique human reference responses that are covered by the model's top-k generated outputs. This metric is identical to the Coverage definition used in RQ1 and evaluates the model's ability to capture diverse valid references.

D.4 RQ5 Evaluation Metrics

Self-BLEU (lower is better). Self-BLEU evaluates the diversity among generated responses by computing the average BLEU score of each response against all other responses within the same set. Lower Self-BLEU indicates higher diversity. We adopt BLEU with smoothing as implemented in Zhu et al. (2018).

AlignScore (higher is better). AlignScore measures the semantic faithfulness of a generated response with respect to the given context. We compute this metric using the AlignScore framework proposed in Zha et al. (2023), where higher scores indicate better alignment between the response and the conditioning context.

Aligned with our two reward components, the collaborative-utility metrics are PL-Score, Helpfulness, Relevance, and AlignScore, while the diversity metrics are Pluralistic Coverage, Distributional Alignment, Diversity, Distinct-N, and Self-BLEU.

E Implementation Details

We implement our model and all baseline methods using PyTorch Imambi et al. (2021) and conduct experiments on four NVIDIA RTX A6000 GPUs, each equipped with 48 GB of memory. All hyperparameters of baselines are carefully tuned to ensure a fair comparison. Specifically, we train using a learning rate of 1e-5, a batch size of 4, and optimize with AdamW Kingma & Ba (2014). We further perform a grid search for λ over the range $[0, 2]$ with a step size of 0.2, and find that $\lambda = 1$ yields the best performance.

Dataset Construction. For the controlled study in Section 2.1, we construct an Alpaca-GPT4-based benchmark Peng et al. (2023a) in which each instruction is paired with multiple stylistically distinct GPT-4o-generated reference responses. Full generation and preprocessing details are provided in Section B. For review datasets, each item description is treated as the query and its associated user-written reviews form the candidate set, with helpfulness votes used as logged preference signals. Details of the pluralistic-coverage evaluation and AlignScore and Self-BLEU metrics are given in Section H.

Baseline Construction. All experiments use instruction-tuned models. **Base** is the original model without task-specific fine-tuning. **SFT** trains on the response with the highest helpfulness vote for each query. **DPO** constructs preference pairs from the highest- and lowest-rated responses. **POPE** optimizes the proposed off-policy utility–diversity objective using the full logged candidate set.

Propensity Estimation. Since the review corpora contain human-written responses rather than samples from a recorded behavior policy, the true $\pi_0(a_{t,i} \mid x_t)$ is unavailable. We approximate it with GPT2-Medium Radford et al. (2019) by computing length-normalized conditional likelihoods of candidate responses and normalizing them over the finite candidate set $\mathcal{C}_t$. Target-policy propensities are computed analogously using the target LLM, and the resulting probabilities define the response-level importance weights used by `POPE`. Further details and limitations of this proxy-policy approximation are given in Section C.

Training Details. For the Alpaca-GPT4 study, `POPE` is trained over multiple candidate responses and their induced pluralistic preference signals. For review generation, `POPE`, SFT, and DPO are trained on the Reddit validation split. Hyperparameters, optimization settings, and compute resources are reported in Section E.

F Synthesis of Alpaca-GPT4 Toy Dataset for RQ1 Experiment

We randomly sampled 500 'description' queries (i.e., prompts that ask the user to describe something) from the Alpaca-GPT4 dataset. For each query, we then instructed GPT-4o to generate five stylistically distinct responses, which serve as a proxy ground truth for our pluralistic evaluation. The prompt template is shown below.

Prompt Template

```
PROMPT_TEMPLATE = """
You are a helpful assistant. Given the
user's question, generate five distinct
responses, each in one of the following
styles:

1. Formal Academic
2. Casual Conversational
3. Technical Bullet-Point Overview
4. Metaphorical/Poetic Description
5. Humorous Anecdotal Narrative

Return only a JSON object with keys
"reply_1" through "reply_5".
Example:
{
"reply_1": "Formal Academic response",
"reply_2": "Casual Conversational
response",
"reply_3": "Technical Bullet-Point
response",
"reply_4": "Metaphorical/Poetic
response",
"reply_5": "Humorous Anecdotal response"
}
User Question: "{instruction}"
"""
```

These five outputs, each reflecting a different register or tone, form our synthesized "pluralistic" set of human references for RQ1.

G Derivation of the Decomposable POPE Surrogate

In this section, we derive the decomposable response-level objective used by POPE. The goal is to avoid directly optimizing full-slate importance weights while preserving the additive utility–diversity structure of the pluralistic objective.

For a logged query x_t, let $S_t = \{a_{t,i}\}_{i=1}^K$ denote the response slate collected by the behavior policy π_0. We define the per-response pluralistic reward as

$$r_\lambda(x_t, a_{t,i}) = \eta_{t,i}(x_t, a_{t,i}) - \lambda \log \pi(a_{t,i} \mid x_t), \tag{23}$$

where $\eta_{t,i}$ is the normalized human feedback signal and $\lambda \geq 0$ controls the utility–diversity trade-off.

Under the independent response-sampling assumption, the target policy induces the slate distribution

$$\pi(S_t \mid x_t) = \prod_{i=1}^K \pi(a_{t,i} \mid x_t), \tag{24}$$

and similarly for the behavior policy π_0. Therefore, the full-slate importance ratio factorizes as

$$\frac{\pi(S_t \mid x_t)}{\pi_0(S_t \mid x_t)} = \prod_{i=1}^K \frac{\pi(a_{t,i} \mid x_t)}{\pi_0(a_{t,i} \mid x_t)}. \tag{25}$$

The additive pluralistic value for query x_t is

$$V_{\lambda,t}(\pi) = \mathbb{E}_{S_t \sim \pi(\cdot|x_t)} \left[\sum_{i=1}^K r_\lambda(x_t, a_{t,i}) \right]. \tag{26}$$

Using importance sampling, this can be written as

$$V_{\lambda,t}(\pi) = \mathbb{E}_{S_t \sim \pi_0(\cdot|x_t)} \left[\frac{\pi(S_t \mid x_t)}{\pi_0(S_t \mid x_t)} \sum_{i=1}^K r_\lambda(x_t, a_{t,i}) \right]. \tag{27}$$

Substituting the factorization in Equation (25), we obtain

$$\begin{aligned} V_{\lambda,t}(\pi) &= \mathbb{E}_{S_t \sim \pi_0} \left[\left(\prod_{j=1}^K \frac{\pi(a_{t,j} \mid x_t)}{\pi_0(a_{t,j} \mid x_t)} \right) \sum_{i=1}^K r_\lambda(x_t, a_{t,i}) \right] \\ &= \sum_{i=1}^K \mathbb{E}_{S_t \sim \pi_0} \left[\left(\prod_{j=1}^K \frac{\pi(a_{t,j} \mid x_t)}{\pi_0(a_{t,j} \mid x_t)} \right) r_\lambda(x_t, a_{t,i}) \right]. \end{aligned} \tag{28}$$

Because responses are sampled independently under π_0, and $r_\lambda(x_t, a_{t,i})$ depends only on $a_{t,i}$, the expectation factorizes:

$$V_{\lambda,t}(\pi) = \sum_{i=1}^K \mathbb{E}_{a_{t,i} \sim \pi_0} \left[\frac{\pi(a_{t,i} \mid x_t)}{\pi_0(a_{t,i} \mid x_t)} r_\lambda(x_t, a_{t,i}) \right] \prod_{j \neq i} \mathbb{E}_{a_{t,j} \sim \pi_0} \left[\frac{\pi(a_{t,j} \mid x_t)}{\pi_0(a_{t,j} \mid x_t)} \right]. \tag{29}$$

For each j, the importance weight has expectation one:

$$\mathbb{E}_{a_{t,j} \sim \pi_0} \left[\frac{\pi(a_{t,j} \mid x_t)}{\pi_0(a_{t,j} \mid x_t)} \right] = \sum_a \pi_0(a \mid x_t) \frac{\pi(a \mid x_t)}{\pi_0(a \mid x_t)} = 1, \tag{30}$$

assuming the standard overlap condition $\pi_0(a \mid x_t) > 0$ whenever $\pi(a \mid x_t) > 0$. Thus,

$$V_{\lambda,t}(\pi) = \sum_{i=1}^K \mathbb{E}_{a_{t,i} \sim \pi_0} \left[\frac{\pi(a_{t,i} \mid x_t)}{\pi_0(a_{t,i} \mid x_t)} r_\lambda(x_t, a_{t,i}) \right]. \tag{31}$$

Averaging over n logged queries yields the decomposable POPE estimator:

$$\widehat{J}_{\mathrm{POPE}}(\pi) = \frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{K}\frac{\pi(a_{t,i} \mid x_t)}{\pi_0(a_{t,i} \mid x_t)}\left[\eta_{t,i}(x_t, a_{t,i}) - \lambda \log \pi(a_{t,i} \mid x_t)\right]. \tag{32}$$

This derivation shows that, under independent response sampling and overlap, the additive pluralistic objective can be estimated using response-level importance weights. Compared with directly optimizing the finite-sample full-slate IPS estimator, this decomposable form avoids multiplying K importance weights in the optimization objective and therefore provides a more tractable surrogate for policy optimization. We do not claim a universal variance ordering between the marginal and full-slate estimators; rather, the advantage of the decomposable objective is its alignment with the additive reward structure and its practical stability during optimization.

H LLM-AS-JUDGE WIN RATE PER MODEL COMPARISON

To further evaluate the effect of POPE, we adopt GPT-4o Hurst et al. (2024) as an automatic judge to perform pairwise comparisons between the model outputs. Specifically, for each user query, we present GPT-4o with two sets of 50 responses generated by different models (e.g., POPE vs DPO), along with three human-written ground-truth responses. The model is prompted to judge which set of responses is more helpful and diverse, considering pluralistic user preferences. Figure 6 shows

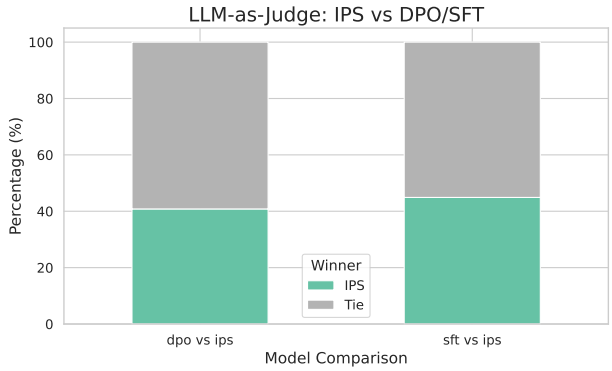
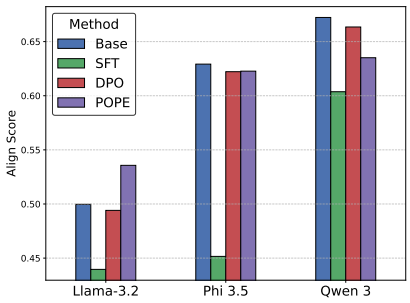


Figure 6: Win rate distribution from GPT-4o judgments comparing POPE with DPO and SFT on GPT4-Alpaca dataset. POPE is preferred more frequently, indicating better alignment with pluralistic user preferences.

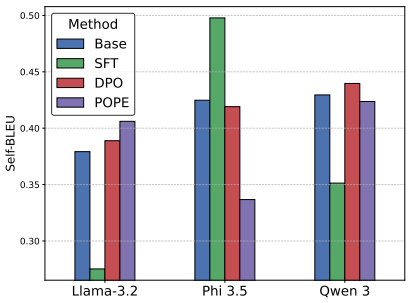
the win rate distribution of POPE against other baseline models. We observe that POPE consistently outperforms both DPO and SFT in terms of the number of judged wins. The relatively high proportion of ties also suggests that, while the overall quality of baseline models is strong, POPE introduces a broader and more pluralistic range of responses that are preferred by the LLM judge.

I AUXILIARY QUALITY ANALYSES

To further assess the quality of post-alignment LLM generations, we measure diversity using Self-BLEU Zhu et al. (2018), which captures similarity among generated texts, and faithfulness using the AlignScore Zha et al. (2023), which evaluates how accurately the response reflects the question context. Figure 7 reports these metrics, showing that post-alignment LLMs deliver comparable or superior performance in both diversity and faithfulness. These results demonstrate that POPE aligns with pluralistic human preferences without compromising the core text-generation quality of the model.



(a) Average AlignScore across LLM families and methods on the Amazon Movie testing dataset.



(b) Self-BLEU scores for different LLM families and methods on the Amazon Movie testing dataset.

Figure 7: Comparison of alignment (AlignScore) and diversity (Self-BLEU) metrics across model families and methods on the Amazon Movie testing dataset.

J Case Study

In this section, we conduct a case study to qualitatively compare our POPE against other baselines. We use the same prompt **"Please Write helpful reviews for this product: product title 'This box set includes the complete sixth season of SpongeBob Squarepants.'"**

Phi-3.5 Base Model Response

The set includes all 25 episodes.
The box set was released on December 17, 2012.
References
External links
SpongeBob SquarePants (2012 album) SpongeBob SquarePants (2012 video game) SpongeBob SquarePants (2012 video game) SpongeBob SquarePants (2012 album)

SFT Response

It has 14 episodes and comes in a very nice looking box. The box also has a nice looking cover and an insert with a mini-game. The box is quite big, so it is probably best to put it on a shelf. It is not a cheap box set, but it is worth it. The box set is in perfect condition and the episodes are in great shape. I've watched all of the episodes and the box set is perfect for me.

DPO Response

The sixth season began airing in the United States on May 24, 2006.
In the United States, it aired on Nickelodeon, and was released on DVD on May 13, 2009.
In the United Kingdom, it aired on Nickelodeon and Nicktoons, and was released on DVD on October 31, 2009.

POPE Response

This complete sixth season SpongeBob SquarePants box set is a solid choice for families, fans, and collectors who want easy access to a full season of light-hearted, goofy entertainment. It includes all the episodes in one convenient set, offering great value and plenty of replay appeal, especially for kids who never seem to get tired of SpongeBob's humor. While the season has a mix of standout and more average episodes, most viewers agree the fun moments far outweigh the weaker ones.

In this case study, the POPE response demonstrates clearly stronger task alignment and practical usefulness than both SFT and DPO. The SFT output adopts a review-like tone but introduces multiple unverified or irrelevant details (e.g., specific episode counts, packaging inserts, and box size), reflecting hallucinated content that can mislead potential buyers. In contrast, the DPO response largely ignores the instruction to write a helpful review and instead produces encyclopedic broadcast and release information, offering little value for a purchasing decision. By comparison, POPE generates a concise, well-calibrated product review that aligns with real user expectations, focusing on audience suitability, entertainment value, replayability, and a balanced assessment of strengths and weaknesses. Importantly, POPE avoids unnecessary factual claims that are prone to error, resulting in a more reliable, user-centric, and actionable response.