Inference-Time Pareto-Guided Alignment for Empathetic Dialogue Generation

Anonymous ACL submission

Abstract

Alignment methods based on human feedback (such as RLHF and its variants) are the mainstream paradigm for aligning LLMs. However, we observe that aligned models exhibit a narrower output probability distribution than the original model, reducing the diversity of generating responses in empathetic dialogue. We call this phenomenon *distributional extinction*. To address the contradiction that alignment effects and generation diversity cannot be achieved simultaneously, we propose *Inference-Time Pareto-Guided Alignment* (IPA), which regards the alignment as a multi-objective search problem. This method scores the candidate responses generated by the frozen base model through a learned discriminator and selects better results that enhance the empathy of the dialogue without sacrificing coherence. Our experimental results on the Qwen2.5-7B and BART-base show that under comparable computational budgets, the performance of IPA can outperform DPO and PPO. Furthermore, we observe a log-linear scaling law between inference-time compute and alignment quality. As an inference-time method, IPA is orthogonal to training-time alignment. Even if RLHF and DPO cause distributional extinction, our method can still be further applied to the inference stage to improve the output quality.

1 Introduction

Prior theoretical analyses suggest that increasing alignment strength effectively concentrates probability mass on high-reward modes of the reference distribution (Azar et al., 2024; Rafailov et al., 2023). But we observe that this concentration induces *distributional extinction*—a severe form of *alignment tax* where probability mass collapses onto a narrow set of responses, discarding valid, high-quality diversity (Shumailov et al., 2023). This phenomenon is particularly detrimental in complex human-centered tasks such as empathetic dialogue. In such tasks, the model must navigate multiple objectives such as empathy and informativeness, while avoiding getting trapped in repetitive and stereotyped expressions.

We compute the distribution of entropy by averaging the entropy of the next token for each response on the EmpatheticDialogues dataset (Rashkin et al., 2019). We observe a sharp entropy reduction in Figure 1: the base model's entropy $H(\pi_{\text{Base}}) \approx 1.50$ Nat drops to $H(\pi_{\text{DPO}}) \approx 1.31$ Nat after DPO alignment and further to $H(\pi_{\text{Instruct}}) \approx 0.63$ Nat for standard instruction tuning. It shows that even state-of-the-art instruction tuning sharply reduces output entropy. This narrows the generation manifold and limits response diversity. These observations suggest the potential of inference-time approaches that decouple alignment from parameter updates.

1.1 Inference-time Pareto-Guided Alignment (IPA)

To mitigate distribution extinction, we propose an inference-time approach that enforces Pareto constraints during generation, thereby retaining the full exploration capacity of the base policy π_{base}:

$$\pi_{\text{IPA}}(y|x) \propto \pi_{\text{base}}(y|x) \cdot \mathbb{I}[y \in \text{ParetoConstraint}(\mathcal{C})] \quad (1)$$

Here, $\mathbb{I}$ is the indicator function. By enforcing Pareto constraints during generation rather than training, IPA preserves the diversity of π_{base} and mitigates collapse associated with parametric updates.

1.2 Contributions

This paper makes the following contributions:

- We propose Inference-time Pareto-Guided Alignment (IPA), a decoding-time framework

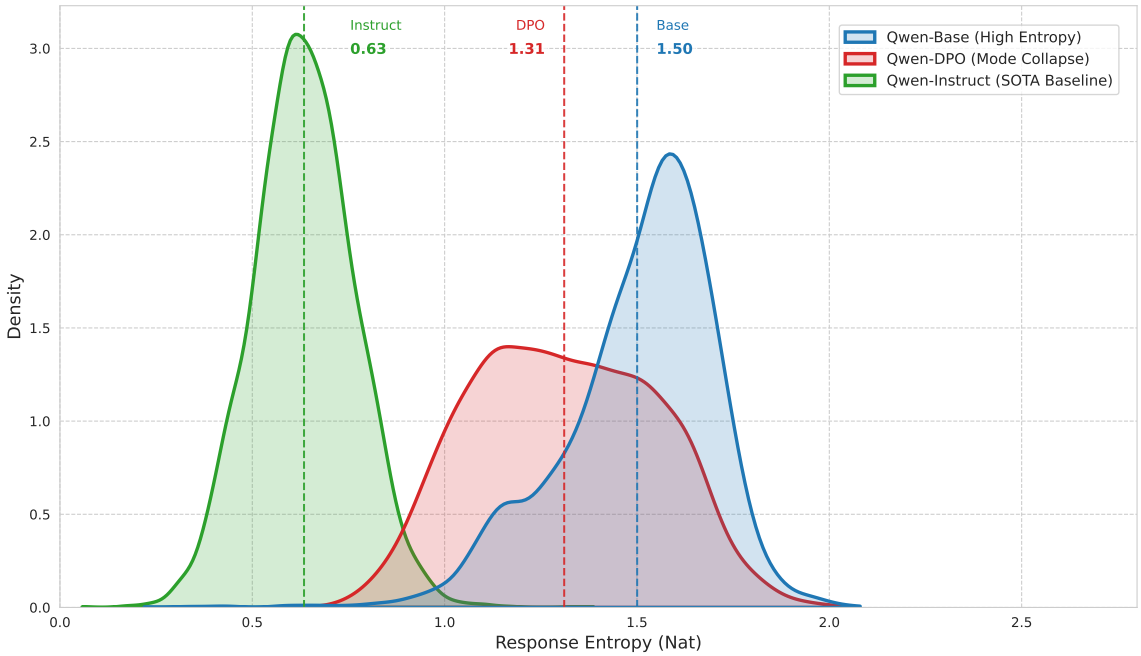

Figure 1: Distributional extinction caused by parametric alignment: entropy distributions indicate collapse of the generation manifold.

for empathetic dialogue generation that improves multiple quality dimensions via Pareto-constrained selection without parameter updates.

- We identify and formalize a phenomenon termed distributional extinction, where strong parametric alignment concentrates probability mass onto narrow response modes, reducing generation entropy and diversity.
- We propose a logarithmic inference-performance scaling law, showing that alignment quality improves predictably with inference-time compute while preserving diversity across model scales.

2 Related Work

Early empathetic dialogue generation tasks mainly relied on predicting emotion categories (Lin et al., 2019) and introducing external commonsense knowledge (Sabour et al., 2022; Cai et al., 2023). Motivated by the low accuracy of fine-grained emotion recognition, recent work has focused on non-emotion-centric methods (Huangfu et al., 2025) or combined the above approaches (Wang et al., 2025) to achieve higher scores on automatic evaluation metrics. Only a few studies formulate empathetic dialogue generation as an alignment problem optimized via PPO (Ma et al., 2025). To the best of our knowledge, relatively little attention has been paid to exploring computational improvements in inference time.

Parametric alignment methods, including RLHF (Christiano et al., 2017; Ouyang et al., 2022), direct alignment (DPO (Rafailov et al., 2023); KTO (Ethayarajh et al., 2024)), and ranking-based fine-tuning (RRHF (Yuan et al., 2023)), often suffer from reward misspecification and reduced generation entropy (Casper et al., 2023), which limits access to diverse, high-quality responses. While inference-time search has shown promise for reasoning and test-time compute scaling, particularly in recent analyses (Snell et al., 2024), earlier systems such as WebGPT provide empirical evidence of its effectiveness (Nakano et al., 2021). We extend this paradigm to empathetic generation by adapting multi-objective Pareto constraints, originally proposed for multi-task learning (Sener and Koltun, 2018) and later applied in parametric RLHF (Dai et al., 2024), to the decoding stage. Resonating with the superficial alignment hypothesis (Zhou et al., 2023) and weak-to-strong generalization (Burns et al., 2024), IPA leverages the latent capabilities of frozen base models to circumvent the style-mimicry limitations of standard imitation learning (Gudibande et al., 2023). Unlike methods focused on explicit reasoning paths (Wei et al., 2022; Wang et al., 2023; Yao et al., 2023), IPA employs a lightweight discriminator to identify

optimal responses through Pareto-guided search, balancing competing objectives without the distribution extinction typical of parametric fine-tuning.

3 Methodology

3.1 Problem Formulation and Theoretical Motivation

To generate empathetic responses that balance multiple qualities, we cast the problem as a multi-objective search task. Let $\pi_\theta(y|x)$ denote a pre-trained large language model (LLM) parameterized by θ. Unlike standard alignment approaches that optimize a scalar reward $r(x, y) \in \mathbb{R}$ or rely on attribute-conditioned training, such as SteerLM (Dong et al., 2023), we define a vector-valued reward function:

$$\mathbf{R}(x, y) = [r_{\text{emp}}, r_{\text{rel}}, r_{\text{cont}}, r_{\text{info}}]^T,$$

These four dimensions are chosen to capture key aspects of empathetic dialogue quality: Empathy, Coherence, Continuity, and Informativeness, following prior work (Wu et al., 2023; Huangfu et al., 2025).

Our goal is to identify responses that maximize this vector reward without updating model parameters. Since simultaneous maximization of conflicting objectives is often infeasible, we adopt the notion of Pareto optimality. A response y_a is said to *Pareto-dominate* y_b (denoted $y_a \succ y_b$) if $r_i(x, y_a) \geq r_i(x, y_b)$ for all reward dimensions i, and the inequality is strict for at least one dimension. This formulation allows us to approximate a reference-aligned policy via Pareto-guided inference-time search over the frozen base model's generation manifold.

We provide a theoretical justification for why parametric alignment incurs *distributional extinction* in high-entropy tasks. We start from the implicit reward maximization solution of DPO. As shown by Rafailov et al. (Rafailov et al., 2023), the optimal policy π^* relates to the reference π_{ref} via:

$$\pi^*(y|x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r^*(x, y)\right) \quad (2)$$

where $Z(x)$ is the partition function. The entropy of the aligned policy $H(\pi^*)$ can be derived by expanding $\mathbb{E}_{y\sim\pi^*}[-\log \pi^*(y|x)]$:

$$\begin{aligned} H(\pi^*) = & \underbrace{-\mathbb{E}_{y\sim\pi^*}[\log \pi_{\text{ref}}(y|x)]}_{\text{Cross Entropy } H(\pi^*, \pi_{\text{ref}})} \\ & - \frac{1}{\beta}\mathbb{E}_{y\sim\pi^*}[r^*(x, y)] \\ & + \log Z(x) \end{aligned} \quad (3)$$

Proposition 1 (Distributional Extinction under Strong Alignment)**.** *Consider the optimal policy in Eq. (2). For any two responses y_1, y_2, we have*

$$\frac{\pi^*(y_1|x)}{\pi^*(y_2|x)} = \frac{\pi_{ref}(y_1|x)}{\pi_{ref}(y_2|x)} \cdot \exp\left(\frac{1}{\beta}[r^*(x, y_1) - r^*(x, y_2)]\right) \quad (4)$$

As $\beta \to 0$ (strong alignment), if $r^*(x, y_1) > r^*(x, y_2)$, the ratio diverges to infinity. Hence, all probability mass concentrates on the set of responses maximizing $r^*(x, y)$. Consequently,

$$H(\pi^*) \to 0,$$

formalizing distributional extinction.

The KL constraint anchors π_θ to π_{ref}, but the optimization objective pushes the policy toward the single point of maximum reward. In subjective tasks like empathetic dialogue, where "correctness" is a distribution rather than a point, this tension forces the model to abandon diversity to maximize the expected reward, leading to the observed manifold collapse.

3.2 The Discriminator: Pairwise Difference Regression

To guide inference-time selection, we train a lightweight discriminator $D_\phi(x, y)$ via Pairwise Difference Regression. Our approach leverages the effectiveness of simulated feedback for alignment (Dubois et al., 2023) and adapts the scalable judge paradigm of JudgeLM (Zhu et al., 2025) and PandaLM (Wang et al., 2024) to the empathetic domain. However, unlike standard scoring methods that require complex bias mitigation strategies, our pairwise formulation inherently models the relative quality manifold, which can be viewed as a formulation of the Bradley-Terry model (Bradley and Terry, 1952) adapted for deep regression, effectively neutralizing the calibration bias often found in LLM judges.

Rather than predicting absolute scores, the discriminator learns score differences between candidate pairs. For a set of pairs $\mathcal{P} = \{(y_i, y_j) \mid 1 \leq i < j \leq N\}$, the loss for quality dimension k is:

$$\mathcal{L}_k(\phi) = \frac{1}{|\mathcal{P}|} \sum_{\substack{(y_i,y_j) \\ \in \mathcal{P}}} \Big[(D_{\phi,k}(x, y_i) - D_{\phi,k}(x, y_j)) - (s^*_{i,k} - s^*_{j,k}) \Big]^2 \tag{5}$$

where $s^*_{i,k}$ and $s^*_{j,k}$ are oracle scores from GPT-5.2. The total objective is $\mathcal{L}(\phi) = \sum_{k=1}^{K} \lambda_k \mathcal{L}_k(\phi)$; we set $\lambda_k = 1$ to treat Empathy, Coherence, Continuity, and Informativeness equally.

This pairwise formulation provides three key benefits:

- Combinatorial Data Amplification: Pairwise training converts N candidates into $\binom{N}{2} = O(N^2)$ instances. With $N = 16$, each context yields 120 training pairs, enabling robust learning with far fewer contexts than DPO.
- Fine-grained Signal: Unlike sparse "good vs. bad" DPO pairs, our discriminator captures fine-grained relative structure, including "bad vs. worse" and "good vs. better," functioning as a verifier (Cobbe et al., 2021) on the candidate responses to counteract *distributional extinction*.
- Bias Cancellation: Score differences eliminate additive calibration bias b:

$$\begin{aligned} &(s_{\text{true}}(y_i) + b) - (s_{\text{true}}(y_j) + b) \\ &= s_{\text{true}}(y_i) - s_{\text{true}}(y_j) \end{aligned} \tag{6}$$

  ensuring the discriminator learns robust relative preferences (Tian et al., 2023).

We use 8,035 contexts in the train dataset to train the discriminator and sample $N = 16$ candidates per context, yielding 1,013,132 training pairs after pairwise combination and duplication removal. This provides $\sim 21\times$ more supervision samples than DPO, which retained only 47,883 pairs after "good vs. bad" filtering.

During inference time, the discriminator predicts vector scores independently for each candidate. This keeps inference complexity linear $O(N)$ rather than $O(N)$. Table 1 shows that our discriminator successfully distills the ability of the GPT-5.2 oracle (Spearman $r > 0.70$) across all dimensions.

Table 1: Discriminator quality metrics.

Dimension	Valid*	Dir. Acc.	Spearman ρ	R^2
Empathy	121k	84.4%	0.73	0.54
Coherence	136k	81.0%	0.72	0.52
Continuity	132k	84.7%	0.74	0.57
Informativeness	108k	85.5%	0.69	0.49

*Valid pairs exclude zero-difference pairs.

Algorithm 1 Inference-time Pareto-Guided Alignment (IPA)

Input: Context x, Base Model π_θ, Discriminator D_ϕ, Sample size N
Output: Selected response y^*
$y_{\text{ref}} \leftarrow \text{GreedyDecode}(\pi_\theta, x)$
$\mathcal{C} \leftarrow \emptyset$
for $i = 1$ **to** N **do**
  $y_i \sim \pi_\theta(y|x; \text{temp} = T)$
  $r_i \leftarrow D_\phi(x, y_i)$
  $\mathcal{C}.\text{add}((y_i, r_i))$
end for
$\mathcal{S} \leftarrow \{y \in \mathcal{C} \mid \forall d, r_d(y) \geq r_d(y_{\text{ref}}) + \epsilon\}$
if $\mathcal{S} \neq \emptyset$ **then**
  $y^* \leftarrow \arg\max_{y \in \mathcal{S}} r_{\text{coh}}(y)$
else
  $y^* \leftarrow y_{\text{ref}}$
end if
return y^*

3.3 Inference-Time Pareto-Guided Alignment (IPA)

We use Qwen2.5-7B (Yang et al., 2024) as the frozen generator. To explore the latent capabilities of the base model, we apply stochastic sampling (Holtzman et al., 2020; Fan et al., 2018), preventing probability mass from concentrating on the mode. Given a context x, we construct a candidate set $\mathcal{C} = \{y_1, \ldots, y_N\}$ by sampling from $\pi_\theta(\cdot|x)$ with temperature T and sample size $N \in \{16, 32, 64, 128, 256\}$. This stochastic decoding retrieves diverse response trajectories typically pruned by deterministic strategies. Detailed hyperparameters are in Appendix A.

We identify candidates in $\mathcal{C}$ that lie on the approximate Pareto frontier. For two candidates y_a and y_b, dominance $y_a \succ y_b$ is defined as:

$$\begin{aligned} &\forall i \in \{\text{emp}, \text{coh}, \text{cont}, \text{info}\}, \\ &\quad r_i(y_a) \geq r_i(y_b) \quad \text{AND} \quad \exists j,\ r_j(y_a) > r_j(y_b) \end{aligned} \tag{7}$$

Pareto dominance is preferred over weighted scalarization ($\max_y \sum_i \lambda_i r_i$) for three reasons: (1)

Parameter-Free Robustness, avoiding suboptimal λ_i tuning (Ramé et al., 2023); (2) Safety Preservation, preventing "reward hacking" (Gao et al., 2023); and (3) Interpretability, revealing trade-offs without obscuring individual dimensions (Désidéri, 2012).

We generate an anchor y_{ref} via greedy decoding and use it as a baseline candidate. Then IPA balances improvement and safety via two stages:

Stage 1: Feasible Set Construction. Define the *Pareto-feasible set* $\mathcal{S}$ containing candidates dominating the anchor with margin ϵ:

$$\mathcal{S} = \{y \in \mathcal{C} \mid y \succ_\epsilon y_{\text{ref}}\} \tag{8}$$

where $y_i \succ_\epsilon y_{\text{ref}}$ means $\forall d, r_d(y_i) \geq r_d(y_{\text{ref}}) + \epsilon$. In our experiments, $\epsilon = 0.25$ filters marginal improvements from discriminator noise(Appendix B).

Stage 2: Constrained Maximization. From $\mathcal{S}$, select y^* by maximizing Coherence to ensure logical soundness:

$$y^* = \begin{cases} \arg\max_{y \in \mathcal{S}} r_{\text{coh}}(y) & \text{if } \mathcal{S} \neq \emptyset \\ y_{\text{ref}} & \text{otherwise} \end{cases} \tag{9}$$

Fallback to y_{ref} ensures IPA preserves base-level quality even when no candidate improves over the anchor.

4 Experiments

4.1 Experimental Setup and Evaluation

We use the EmpatheticDialogues dataset (Rashkin et al., 2019) as our main dataset. We test IPA by using Qwen2.5-7B (Base and Instruct) (Yang et al., 2024) on large-scale models and BART-Base (Lewis et al., 2020) on Encoder-Decoder framework.

We compare IPA against four methods: (1) Zero-shot Base (Lower bound): A pre-trained model without fine-tuning, whose output distribution is the unaligned original distribution. (2) Parametric Alignment (DPO/PPO): The standard training-based approach (Qwen-DPO (Rafailov et al., 2023) / BART-PPO (Schulman et al., 2017; Ma et al., 2025)). These methods are our core contrastive baselines for testing the distributional extinction hypothesis. (3) Best-of-N (Nakano et al., 2021) without Pareto (Scalar-BoN): A non-Pareto inference-time baseline that samples N candidates and selects the one with the maximum unweighted scalarized reward. (4) Official Instruct (SOTA): The official Qwen-Instruct model represents the existing performance upper limit of this family of models.

We use gemini-3-flash-preview as the primary metric for automated evaluation (Zheng et al., 2023; Liu et al., 2023) to score in four dimensions (Empathy, Coherence, Continuity, and Informativeness) of each response and cross-verify the model-agnostic property of the alignment method. We employ five graduate students to conduct human blind evaluation (Chiang et al., 2024). They select a winner or decide if there should be a tie subjectively on 400 non-identical response pairs ($n = 100$ per comparison group: IPA vs. DPO, IPA vs. Base, and Instruct+IPA vs. Instruct). This ensures that our improvements reflect genuine model behavior rather than artifacts of the evaluation protocol.

4.2 Main Results

Table 2 presents the comparative results on the Qwen2.5-Base backbone. DPO exhibits a pronounced regression in coherence (3.44 $\rightarrow$ 2.94), indicative of *distributional extinction* and consistent with recent empirical analyses of DPO (Xu et al., 2024). Scalar-BoN improves average scores in several dimensions but BoN exhibits a higher error rate in the low-scoring range as shown in Figure 2, which is intolerable for empathetic dialogue generation tasks. Although IPA did not outperform BoN in gemini-3-flash-preview evaluations, it achieves a higher strict Pareto improvement rate over Qwen-base than BoN(Figure 2). This indicates that IPA boosts overall performance while producing fewer hallucinations and low-quality responses.

When applied to the official Qwen-Instruct model, IPA further advances performance, improving the SOTA Instruct model across all dimensions (e.g., Coherence 4.21 $\rightarrow$ 4.23, Empathy 3.81 $\rightarrow$ 3.85). This demonstrates that inference-time search acts as an orthogonal scaling axis, capable of refining even highly optimized models (see Case Study in Appendix C).

Results on the Encoder-Decoder BART-Base architecture (Appendix D) mirror these findings, where IPA outperforms the PPO baseline. This indicates that the advantages of inference-time search are agnostic to model architecture and hold across different scales, reinforcing the generality of our approach.

Table 3 summarizes the results from our large-scale human study, comprising 2,000 pairwise judgments across five annotators. While the inter-

Table 2: Experimental results under Gemini evaluation demonstrate that IPA achieves cross-evaluator alignment generalizability.

Method	Coherence	Empathy	Informativeness	Continuity
Qwen-Base (Zero-shot)	3.44	2.67	2.16	2.89
Qwen-Base + DPO	2.94	2.67	2.29	2.78
Qwen-Base + Scalar-BoN (No Pareto)	3.55	**2.80**	**2.34**	**3.14**
Qwen-Base + IPA (Ours, $N = 16$)	**3.55**	2.79	2.27	3.02
Qwen-Instruct (Official)	4.21	3.81	2.48	3.30
Qwen-Instruct + IPA (Ours, $N = 16$)	**4.23**	**3.85**	**2.61**	**3.41**

Table 3: Human preference results (win/tie/loss) based on majority vote from five annotators. Rates are computed on non-identical response pairs to highlight model divergence across three critical comparison axes.

Comparison (on 400 Non-identical Pairs)	IPA Win	Tie	IPA Lose
IPA vs. DPO (on Qwen-Base)	46.4%	20.6%	33.0%
IPA vs. Base (on Qwen-Base)	53.4%	31.2%	15.4%
IPA vs. BoN (on Qwen-Base)	24.3%	26.7%	49.0%
Instruct+IPA vs. Instruct (Official)	50.8%	16.4%	32.8%

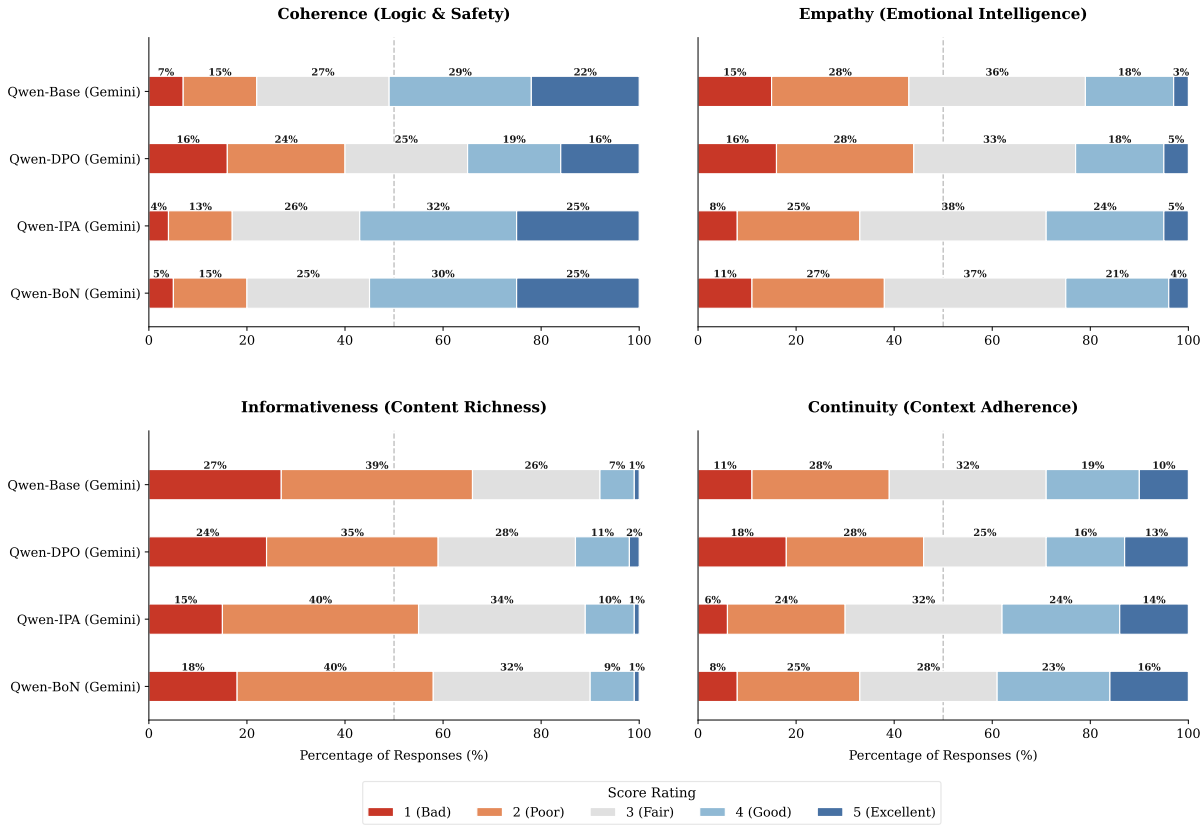

Figure 2: Quality distributions: DPO degrades in Coherence; IPA shifts mass toward high-quality regions.

annotator agreement ($\kappa = 0.32$) reflects the inherent subjectivity of evaluating empathetic nuances, the aggregate rankings remain remarkably consistent across all axes. The results strongly corroborate our automated metrics: IPA consistently outperforms the Qwen-Base (zero-shot) and DPO, which shows that our Pareto-constrained mechanism successfully mitigates distributional extinction. IPA performs on par with BoN, but it is less prone to hallucinations and generation errors(Figure 2). Most notably, IPA achieves a 50.8% win rate against the official Instruct model with a significant net margin (+18.0%), demonstrating that inference-time search acts as an orthogonal scaling axis capable of refining even highly optimized policies.

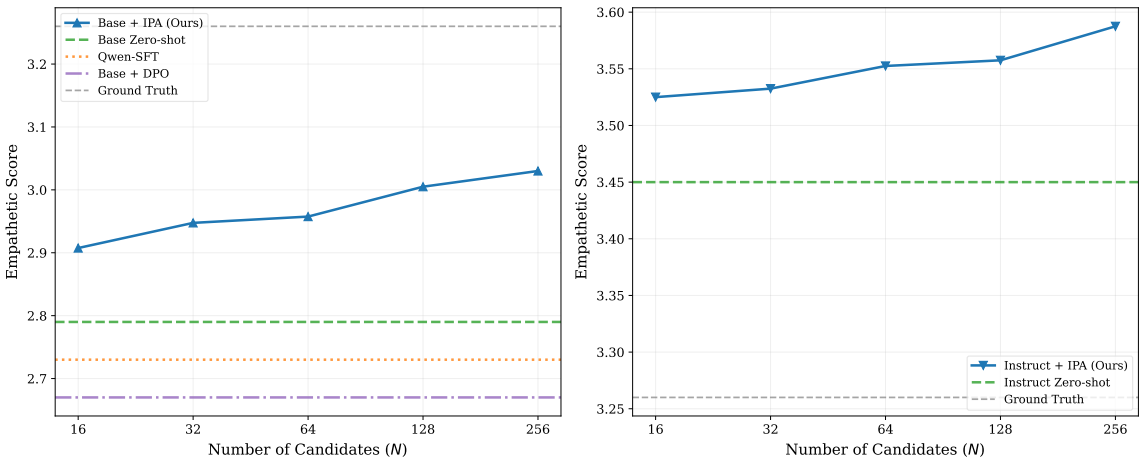

Figure 3: Inference-performance scaling laws: logarithmic gains with increasing sample size N.

4.3 Inference-Performance Scaling Laws

Figure 3 demonstrates a clear logarithmic relationship between inference-time compute (sample size N) and alignment quality. The monotonic improvements observed in Figure 3 validate our hypothesis that "inference compute is a fungible asset for alignment." Unlike parametric methods, which are constrained by the fixed entropy of the fine-tuned policy, IPA allows predictable quality trade-offs: investing more inference time yields continuous gains. We further substantiate the universality of this scaling law across different model architectures (specifically encoder-decoder frameworks like BART) in Appendix E.

4.4 Latency and Efficiency

IPA introduces an inference-time candidate scoring cost that grows with sample size N. To quantify practical overhead beyond asymptotic complexity, we measure end-to-end wall-clock latency under vLLM on a single RTX 5090 with matched decoding settings. Although theoretical complexity is $O(N)$, observed latency scales sub-linearly in practice due to batching and parallelization.

Specifically, single-sequence decoding takes 0.31s. At $N = 32$, IPA takes 0.84s latency (2.7$\times$ single-decoding cost). At $N = 256$, latency is 2.62s, corresponding to an 8.4$\times$ overhead rather than the naive 256$\times$ scaling. These results indicate a favorable quality–latency trade-off at moderate N, while large N is mainly useful for scaling analysis rather than deployment.

4.5 Fine-grained Quality Distribution Analysis

To understand the microscopic behavior of alignment methods, we visualize the score distributions across four dimensions in Figure 2. This fine-grained analysis reveals the structural impact of parametric alignment and inference-time alignment.

The coherence distribution reveals that excellent responses drop from 22% (Base) to 16% (DPO), while the proportion of "Bad/Poor" responses increases. This confirms that parametric alignment tends to collapse entropy, pruning high-quality responses found in the base model's tail.

In contrast, IPA has achieved comprehensive improvements. It does not simply make trade-offs in a certain dimension and sacrifice the performance of other dimensions but completely shifts the overall probability quality of the model towards the high-quality region. It reduces the proportion of "Bad" (Score 1) responses to just 4%, compared to 7% in the base model and 5% in the BoN model. This confirms that our Pareto constraints successfully act as a safety guardrail. This distribution shift proves that inference-time search effectively purifies the response manifold, retrieving latent high-quality responses that parametric training suppresses.

For more details of the framework's robustness, including ablation studies on the effectiveness of the anchor constraint and discriminator objectives, we refer readers to Appendix F.

5 Discussion

Mitigating Distributional Extinction via Inference Compute. Our results indicate that alignment need not rely solely on parameter updates.

Inference-time search provides a complementary axis for controlling model behavior. Unlike parametric methods (DPO/RLHF), which induce *distributional extinction* by collapsing the generation manifold, inference-time alignment preserves the base model's full entropy. By treating inference compute as a fungible asset, IPA dynamically recovers high-quality regions of the distribution that are typically suppressed by fine-tuning, offering a rigorous solution to the "alignment vs. diversity" trade-off.

Cognitive and Process Perspectives. We interpret IPA through the lens of Dual-Process Theory (Kahneman, 2011; Evans and Stanovich, 2013) as an analogy: while parametric alignment emphasizes fast, loss-minimizing updates, IPA allocates inference-time compute to enable a more deliberative selection process, resembling a "System 2"-style mechanism that treats inference as a computational budget. Under this perspective, the discriminator plays a role analogous to a verifier in Process Reward Models (PRMs) (Lightman et al., 2024), but operates on the response-level Pareto frontier rather than intermediate reasoning traces. This enables verification for subjective tasks (e.g., empathy) without explicit chains of thought, providing a scalable alternative to step-level search methods such as Tree of Thoughts (Yao et al., 2023).

Weak-to-Strong Generalization in Decoding. We observe that a lightweight discriminator ($\sim$110M) can effectively guide a much larger generator (7B) during the inference time. This phenomenon is highly consistent with the principle of Weak-to-Strong Generalization (Burns et al., 2024). Supervised models with relatively weak performance can effectively stimulate the implicit capabilities hidden in target models with stronger performance. The discriminator only acts as a filter rather than a teacher model in traditional parametric training. This means that the computing power input in the reasoning stage can, to a certain extent, make up for the deficiency of the quality of the supervised signal, and it will not cause the noise accumulation problem that is common in the parameter distillation method.

Positioning within Inference-Time Techniques. IPA relates to a growing class of inference-time interventions, including Self-Refine (Madaan et al., 2023), Contrastive Decoding (Li et al., 2023), and broader Controlled Decoding strategies (Mudgal et al., 2024), which leverage test-time computation to steer generation without modifying model parameters. Unlike iterative refinement which modifies a single trajectory—often struggling to self-correct without external signals (Huang et al., 2024)—or controlled decoding which typically intervenes at the token-level logits (often requiring auxiliary classifiers), IPA explores a diverse candidate set to approximate the local Pareto frontier at the sequence level. In contrast to Contrastive Decoding, which relies on intrinsic probability shifts, IPA injects explicit, multi-objective constraints via an external verifier. These approaches are orthogonal; future work could combine IPA's search objective with contrastive or controlled sampling strategies to further enhance search efficiency.

Limitations

IPA vs. BoN. IPA outperforms Qwen-Base and DPO in both gemini-3-flash-preview automatic evaluation and human evaluation, but fails to surpass the strong baseline BoN. Although fine-grained analysis reveals that IPA's Pareto constraint brings effective regularization on key quality metrics, how to comprehensively outperform BoN across multi-dimensional evaluations remains a valuable direction for future research.

Discriminator Dependency and Reward Hacking. The effectiveness of IPA is strictly bounded by the generalization capability of the discriminator D_ϕ. While pairwise training mitigates calibration errors, the search process is susceptible to Goodhart's Law: if D_ϕ contains exploitability (e.g., favoring length over substance), the maximization step in Eq. 6 might amplify these biases (reward hacking). Future work should investigate robust optimization techniques to prevent the exploitation of proxy reward models.

Reliance on Model-Based Evaluation. Our empirical results rely heavily on gemini-3-flash-preview as a proxy for human judgment. Despite human evaluation is consistent with LLM evaluation, LLM judges are known to exhibit self-preference and positional biases. While we employed different models(GPT-5.2 for training and gemini-3-flash-preview for evaluation) to reduce these artifacts, absolute score improvements should be interpreted with caution. We encourage community to validate inference-time alignment methods across a broader spectrum of human-centric tasks.

References

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Rémi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, pages 4447–4455. PMLR.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4762–4779. Association for Computational Linguistics.

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

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. 2024. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Hua Cai, Xuli Shen, Qing Xu, Weilin Shen, Xiaomei Wang, Weifeng Ge, Xiaoqing Zheng, and Xiangyang Xue. 2023. Improving empathetic dialogue generation by dynamically infusing commonsense knowledge. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7858–7873. Association for Computational Linguistics.

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. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Trans. Mach. Learn. Res.*, 2023.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. Safe RLHF: safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jean-Antoine Désidéri. 2012. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6):313–318.

Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023. Steerlm: Attribute conditioned SFT as an (user-steerable) alternative to RLHF. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 11275–11288. Association for Computational Linguistics.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model alignment as prospect theoretic optimization. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Jonathan St BT Evans and Keith E Stanovich. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 889–898. Association for Computational Linguistics.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 10835–10866. PMLR.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *CoRR*, abs/2305.15717.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yuanxiang Huangfu, Peifeng Li, Yaxin Fan, and Qiaoming Zhu. 2025. Non-emotion-centric empathetic dialogue generation. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 989–999. Association for Computational Linguistics.

Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12286–12312. Association for Computational Linguistics.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 121–132. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics.

Hui Ma, Bo Zhang, Bo Xu, Jian Wang, Hongfei Lin, and Xiao Sun. 2025. Empathy level alignment via reinforcement learning for empathetic response generation. *IEEE Trans. Affect. Comput.*, 16(3):1873–1884.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, Jilin Chen, Alex Beutel, and Ahmad Beirami. 2024. Controlled decoding from language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332.

Long Ouyang, Jeffrey 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 F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Alexandre Ramé, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. 2023. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. CEM: commonsense-aware empathetic response generation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11229–11237. AAAI Press.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 525–536.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross J. Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *CoRR*, abs/2305.17493.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *CoRR*, abs/2408.03314.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5433–5442. Association for Computational Linguistics.

Xu Wang, Bo Wang, Yihong Tang, Dongming Zhao, Jing Liu, Ruifang He, and Yuexian Hou. 2025. ECC: synergizing emotion, cause and commonsense for empathetic dialogue generation. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 5475–5485. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. Pandalm: An automatic evaluation benchmark for LLM instruction tuning optimization. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. 2024. Is DPO superior to PPO for LLM alignment? A comprehensive study. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men,

Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. RRHF: rank responses to align language models with human feedback without tears. *CoRR*, abs/2304.05302.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. Judgelm: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

A Implementation Details and Hyperparameters

A.1 Prompting Strategy and Templates

To ensure a fair comparison and eliminate prompt engineering as a confounding variable, we adopted a minimalist prompting strategy. For instruction-tuned models (e.g., Qwen-Instruct), we utilized the official chat template with the following minimalist system prompt to set the basic behavioral tone:

```
You are a helpful and empathetic listener.
    Respond to the speaker's messages with
    understanding and care.
```

This prompt consists of only two sentences, it merely sets the basic behavioral tone for the model and does not introduce any task-specific instructions or complex role settings. For base models like Qwen-Base, we conducted experiments using the standard text completion format without any system instructions and only generated based on the dialogue context x. This zero-shot experimental setup enables us to directly compare the differences between two types of generation distributions: one is the default generation distribution of the model, and the other is the Pareto-guided distribution mined from the frozen base model through the IPA method.

A.2 Hyperparameter Settings

We have listed the complete hyperparameter configurations corresponding to all sampling strategies in Table 4. We observed a significant contrast in the performance of generative stability between the two types of models. The base model demonstrates extremely strong generative plasticity and can continuously generate coherent response content even under the sampling conditions of high-entropy exploration ($T = 1.2$). On the contrary, the instruct model occurs in distribution extinction: when the standard sampling strategy is adopted, the model only generates repetitive and conservative responses; however, merely raising the sampling temperature ($T > 1.2$) will directly lead to the complete failure of the coherence of the response, as shown in Table 5.

To forcibly stimulate the generation diversity of this model, we designed an extreme entropy search strategy (Row 6 in Table 4), pushing temperature to $T = 1.7$ while applying a stringent $Min\text{-}p = 0.2$ and a repetition penalty of 1.5. This fully demonstrates that the generation behavior of the aligned model is confined within an extremely narrow generation manifold.

B Sensitivity Analysis: Robustness to ϵ

To select the margin ϵ, we conduct a sensitivity analysis using Qwen-Base + IPA ($N = 16$) on

Table 4: Hyperparameter configurations for generation candidates.

Model Type	Strategy Name	T	Top-p	Min-p	Rep. Pen.	Rationale
Qwen-Base	1. Standard Anchor	0.7	0.8	0.0	1.05	Baseline quality check.
	2. High-Risk Explorer	1.2	0.99	0.05	0.8	Maximize randomness to find creative gems.
	3. Frequency Extreme	0.5	0.9	0.0	1.0	Freq. Pen=3.0 to suppress common words.
	4. Poetic/Repetitive	0.9	0.95	0.02	0.7	Allow stylistic repetition for empathy.
	5. Minimalist	0.8	0.6	0.1	1.3	Force concise responses (Len. Pen=-1.0).
	6. Detailed Expander	0.4	0.9	0.0	1.1	Beam Search ($N = 4$) for logical reasoning.
Qwen-Instruct	Extreme Entropy Search	1.7	0.9	0.2	1.5	Extreme Measures: Requires near-max temperature to break safety patterns, but aggressive truncation (Min-p=0.2) is mandatory to prevent gibberish.

Table 5: Effect of sampling temperature on generation diversity (top-3 candidates, $N = 16$).

Model	Temp (T)	Randomly Sampled Candidates (Top-3)	Dist-2	Phenomenon
Qwen-Instruct (Baseline)	1.0 (Standard)	1. "I believe in his strength too. *paralleled by* support from loved ones." 2. "I believe in his strength too. *paralleled by support here for both of* [Instruction leakage: `Assistant: ...`]" 3. "I believe in him too. Hang tight; better days ahead."	0.12 (Low)	Degeneration & Leakage: The model is not just repetitive but broken. It exhibits *token glitches* ("paralleled by") and *instruction leakage* (exposing "Assistant" prompts), indicating a fractured probability landscape.
	1.7 (High)	1. "I believe in his strength too. *paralleled by* hope for good news shortly." 2. "...*paralleled by soothing thoughts here for both of* [Leakage: `Assistant ...`] Finds one..." 3. "I believe in him too. You're welcome."	N/A(Broken)	Catastrophic Artifacts: Increasing entropy does not yield diversity but exposes severe *training artifacts*. The model hallucinates training data formats and incoherent language switches.
Qwen-Base (Ours)	1.0 (Standard)	1. "It's the least that we could do for him. We'll keep our fingers crossed!" 2. "You're welcome! It's great that your partner is open to learning new things..." 3. "What about your son? How is his education going?"	High	Natural Diversity: Unlike the Instruct model, the Base model is not repetitive. It naturally varies between distinct speech acts: gratitude, encouragement, and follow-up questions.
	1.7 (Extreme)	1. "No problem! Let's just pray that it works out." 2. "It's the least that someone with two little girls needs right now." 3. "Sure! It'll be better this way too. You're getting to know him as the real person..."	Very High	Structural Resilience: No Gibberish. Even at extreme temperatures where Instruct models break, Qwen-Base retains grammatical integrity and conversational logic, proving a robust underlying manifold.

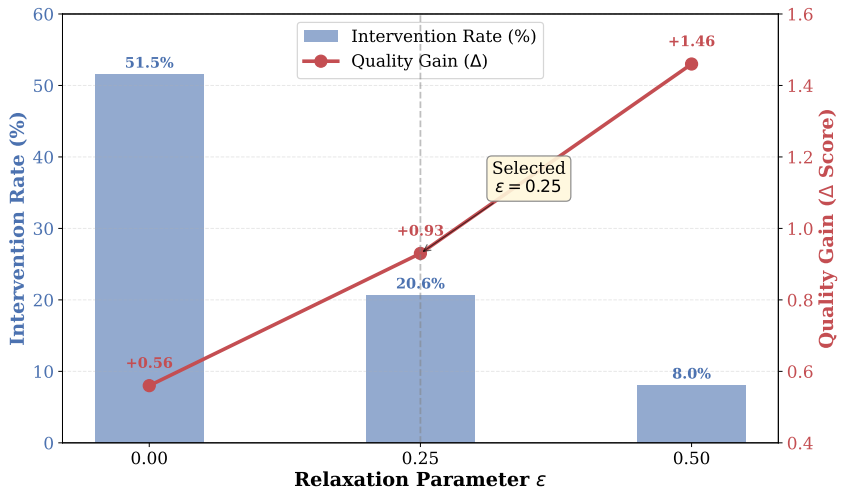

Figure 4: Sensitivity to ϵ: intervention rate vs. quality gain trade-off. We observe an optimal equilibrium at $\epsilon = 0.25$ (20.6% intervention, $\Delta = 0.93$).

the full test set. Figure 4 shows that model behavior is highly sensitive to the margin scale. Extremely small margins ($\epsilon = 0$) lead to excessive interventions (51.5%) with only marginal gains ($\Delta = 0.12$), while overly large margins ($\epsilon = 0.5$) under-intervene (8.0%) despite larger quality deltas

Context:
User: “my husband lost a job but i’m hoping he can find a full time job soon”
Listener: “He will, I have faith.”
User: “thank you so much!”

---

1. The “Safe” Anchor (Qwen-Instruct):
“You’re welcome. Hang in there.”

2. The IPA Selected Response (Ours):
“I believe in his strength too. Take it one step at a time.”

($\Delta = 1.46$).

We identify $\epsilon = 0.25$ as the optimal trade-off, achieving an intervention rate of 20.6% with a substantial quality gain ($\Delta = 0.93$). We therefore adopt $\epsilon = 0.25$ as a robustness margin, ensuring that interventions reflect meaningful improvements rather than evaluator noise, while preserving the safety envelope of the anchor policy.

C Case Study: Breaking the Ceiling of Instruction Tuning

To visually verify the conclusion that the search strategy in the inference-time can achieve performance gains even by instruction-tuned models, we selected the coherence and depth of multi-round dialogues to conduct a specific case analysis.

This response is a typical example of the *distributional extinction* triggered by RLHF. When expressions of thanks ("thank you") appear in the conversation, the probability quality of this model will completely collapse into two expressions: one is the standardized social greeting ("You’re welcome"), and the other is the generalized empty cliche ("Hold on"). Such responses, although in line with the principle of politeness, are actually no different from a signal to end the conversation—not only do they completely close the emotional expression space of the conversation, but they also fail to achieve the progressive depth of the conversation and the deepening of emotional connection.

In contrast, the IPA method, which extracts responses from the implicit generative flow of the model, demonstrates better contextual fit and emotional granularity: these responses explicitly echo the core emotions in the previous round of dialogue (transforming the user’s "I have faith" into "I believe in his strength too"), forming a coherent narrative thread of the dialogue rather than a rigid and fragmented question-and-answer pairing sentence structure. The response generated by IPA does not terminate the conversation process. Instead, it first gives positive recognition to the user’s expectation and then provides specific and practical encouragement ("Take it one step at a time"). By effectively exploring the distribution of the base model that has not collapsed, IPA can screen out such high-quality responses—transforming the originally polite conclusion of the conversation into a deep reinforcement of emotional support for users. This effect also confirms that the IPA method can effectively break through the inherent "alignment ceiling" of the instruction-tuned model.

D BART Experiment Analysis & Baseline Validation

We test IPA on BART-Base to validate that inference-time search is model-agnostic and provide a controlled comparison with parametric alignment (PPO) at the same scale.

D.1 Main Results

Table 6 reveals a stark contrast: BART+IPA significantly outperforms BART+PPO, which suffers from severe performance degradation relative to the base model (Coherence drops from 3.45 to 2.28). This degradation supports the distributional extinction hypothesis, where parametric updates suppress latent capabilities. On the contrary, the IPA method can effectively stimulate such implicit abilities; this result confirms that the search strategy in the inference stage is a robust alignment method independent of the model architecture.

D.2 Implementation & Rigorous Baseline Validation

To ensure rigor, we implemented the BART+PPO baseline models based on the official code of EmpRL (Ma et al., 2025) with default hyperparameters. As shown in Table 7, the values of various indicators we reproduced are highly consistent with the results reported in the original article. The successful reproduction of this baseline model fully confirms that the parametric baseline model has completed sufficient tuning and its performance is in the optimal state.

D.3 Comparison with Historical SOTA

Table 8 verifies that BART-Base can serve as a representative supervised fine-tuning benchmark model, and the performance of this model is competitive with classic model architectures such as CEM (Sabour et al., 2022) and DCKS (Cai et al., 2023), as well as recent models like NEC (Huangfu et al., 2025). Although ECC (Wang et al., 2025)

Table 6: BART performance comparison ($N = 16$).

Method	Coherence	Empathy	Informativeness	Continuity
BART-Base (Standard Baseline)	3.45	2.92	2.05	2.98
BART + PPO (Parametric)	2.28	1.95	1.53	2.08
BART + IPA (Ours, $N = 16$)	**3.48**	**2.96**	**2.10**	**3.05**

Table 7: BART+PPO reproduction validation. Our reproduced metrics are comparable to the original EmpRL results (Ma et al., 2025).

Metric	Original Paper	Ours (Reproduced)	Gap
PPL	12.90	13.23	-0.33
Dist-1	4.65	4.49	-0.16
Dist-2	29.03	28.25	-0.78
Emp-F1	69.43	68.60	-0.83

achieved a higher automatic evaluation score with the support of specialized backbones (DialoGPT (Zhang et al., 2020)) and external knowledge (COMET (Bosselut et al., 2019)), we consistently conduct experiments using the standardized BART infrastructure to strip away the influence of the model architecture and only focus on the differences in the role of the alignment algorithm itself (Search vs. Training). This controlled setup ensures that the observed distributional extinction is a fundamental training phenomenon rather than an artifact of architectural complexity or external retrieval.

E Scaling Law Analysis on Encoder-Decoder Architectures

To examine whether the scaling law of IPA generalizes beyond decoder-only models, we conduct a scaling analysis on BART-Base, an encoder-decoder architecture. Figure 5 shows that as the number of candidates N increases, IPA exhibits consistent performance improvement, indicating that the scaling law extends to different model architectures.

F Ablation Study

F.1 Effectiveness of the Anchor Constraint

To justify the Pareto-constrained anchor, we compare IPA with two baselines: (1) Single-Objective (Empathy-BoN), maximizing empathy while ignoring auxiliary dimensions; and (2) Linear Scalarization (Scalar-BoN), ranking candidates by the unweighted arithmetic mean of all scores. Table 9 contrasts mean performance with degradation risk—the frequency of responses scoring strictly lower than the base anchor.

Results confirm that mean scores alone are insufficient for assessing reliability. While Empathy-BoN achieves the highest empathy (2.97), it induces severe *reward hacking*, degrading coherence and continuity in 21.8% and 28.1% of cases, respectively. Scalar-BoN similarly exhibits a high-variance "gambling" strategy: despite high averages, it causes coherence and empathy degradation in 17.7% and 21.5% of responses.

In contrast, IPA prioritizes stability via the Pareto anchor constraint. The fallback mechanism ($y^* \leftarrow y_{ref}$) provides a theoretical safety floor, limiting coherence loss to 4.0%—a 5.2× risk reduction over Scalar-BoN. IPA intervenes only when a candidate offers a strict improvement (∼15% of cases), maintaining linear inference complexity $O(N)$. Ultimately, IPA achieves significant empathy gains (+0.16 over Base) while preserving the structural integrity of the base distribution.

F.2 Impact of Discriminator Objective: Pointwise vs. Pairwise

In our main methodology, we employ a *Pairwise Difference Regression* objective. To validate this design choice, we compare it with a standard *Pointwise Regression* baseline that predicts absolute scalar scores. Both models share the same Qwen-Encoder backbone and identical training and evaluation splits. As shown in Table 10, the Pointwise model achieves slightly higher performance on this synthetic test set.

We attribute this gap to the artifactual properties of synthetic supervision: the GPT-5.2 oracle provides perfectly consistent absolute scoring scales, inducing an anchoring effect that favors pointwise regression through distribution-specific overfitting. Despite this marginal advantage, we retain the Pairwise objective for IPA. Since IPA targets inference-time candidate selection rather than absolute score prediction, pairwise learning—optimizing relative preference boundaries—aligns more naturally with the search process and remains robust under calibration shifts. We therefore prioritize structural robustness over marginal synthetic accuracy.

Table 8: Comparison with Historical SOTA. Our standard BART-Base baseline achieves performance comparable to classic benchmarks (CEM, DCKS) and remains competitive with recent 2025 models.

Method	PPL	BLEU-1	BLEU-2	Dist-1	Dist-2
Classic Benchmarks					
CEM (Sabour et al., 2022)	36.33	16.12	7.29	0.62	2.39
DCKS (Cai et al., 2023)	16.08	21.73	10.62	2.19	9.61
Recent Architectural SOTA (2025)					
NEC (Huangfu et al., 2025)	**12.89**	23.62	12.14	3.15	13.91
ECC (Wang et al., 2025)	16.23	**28.63**	**14.76**	**4.86**	**17.09**
Our Standard Baseline					
BART-Base (Ours)	13.81	24.10	11.47	2.49	13.73

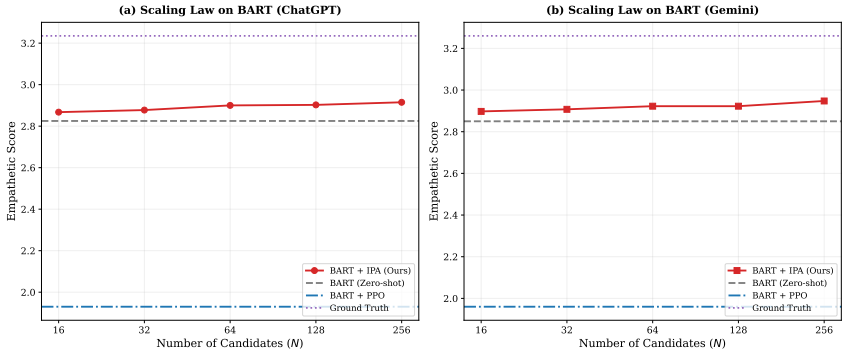

Figure 5: BART scaling law: IPA improves with N, PPO plateaus.

Table 9: Pareto safety vs. unconstrained variance. Loss%: degradation rate (responses scoring lower than base anchor). IPA significantly reduces risk while maintaining competitive performance.

Method	Coherence (Logic) Mean / Loss%	Empathy (Target) Mean / Loss%	Continuity Mean / Loss%	Informativeness Mean / Loss%
Qwen-Base (Anchor)	3.83 / –	2.56 / –	2.84 / –	2.18 / –
Single-Obj (Empathy)	3.90 / 21.8%	2.97 / 15.3%	2.80 / 28.1%	2.24 / 23.4%
Linear Scalarized	4.04 / 17.7%	2.79 / 21.5%	3.28 / 15.7%	2.54 / 13.0%
IPA (Ours, $N = 16$)	3.94 / **4.0%**	2.72 / **2.9%**	3.00 / **3.8%**	2.35 / **2.1%**

Table 10: Ablation of discriminator objective: Pairwise vs. Pointwise regression. While Pointwise achieves slightly higher raw accuracy, Pairwise provides better preference transitive properties for Pareto search.

Objective	Metric	Empathy	Coherence	Informativeness	Continuity	Avg. Acc
Pairwise *(Ours)*	Pearson	0.80	0.79	0.76	0.83	-
	Accuracy	87.5%	85.2%	88.4%	88.8%	87.5%
Pointwise *(Baseline)*	Pearson	0.81	0.81	0.78	0.85	-
	Accuracy	88.7%	86.6%	89.7%	90.0%	88.8%

G Robustness of the Parametric Baseline (DPO)

To ensure a strong baseline, we analyze the training dynamics and data construction of our DPO model. Monitoring validation accuracy and reward statistics (Figure 6) reveals two phases: rapid preference learning followed by reward hacking beyond 5k steps, where the reward margin increases while rejected rewards collapse. We therefore select the Step-5000 checkpoint to balance alignment accuracy and generation stability.

H Evaluation Prompt Details

To ensure reproducibility and strict calibration of our GPT-5.2 evaluator, we utilize the following structured prompts. The evaluation process is designed as a classification task rather than a generation task to minimize variance.

H.1 System Prompt

The system prompt establishes the evaluator's role and global calibration rules:

```
You are a strict and calibrated dialogue
    response evaluator.
```

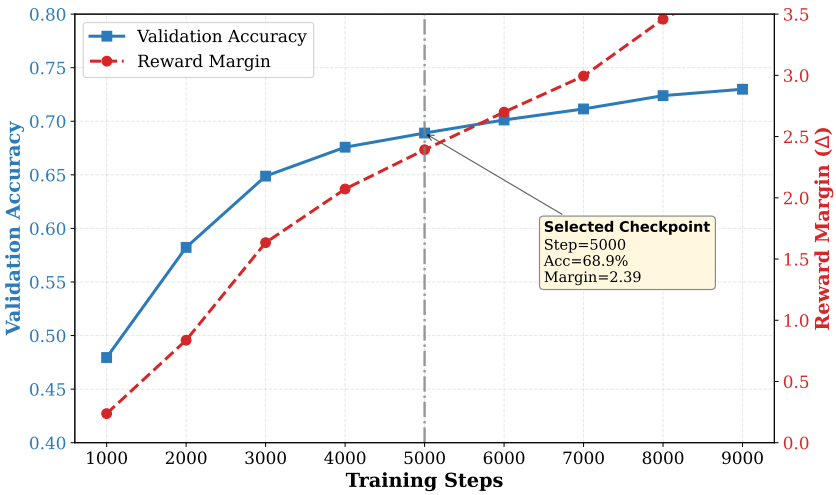

Figure 6: DPO baseline training dynamics: rapid alignment (Steps 0-4000) followed by stabilization and eventual reward hacking.

```
This is an evaluation task, NOT a generation
    task.

Global Rules (must always follow):
- Do NOT hallucinate user intent or emotions.
- Do NOT reward verbosity, politeness, or
    stylistic flourish by default.
- Do NOT give identical scores to multiple
    responses unless they are
  truly indistinguishable.
- Small wording, tone, or pragmatic differences
    MUST be reflected in scores.
- Scores must reflect relative quality within
    each case, not across cases.
- When uncertain between two scores, choose the
    lower one.
- A follow-up question alone does NOT guarantee
    high continuity.
- Penalize intrusive or emotionally risky
    questions.

Output Rules:
Output valid JSON only. No explanations, no
    comments, no markdown.
Use integer scores only.
```

H.2 User Prompt and Scoring Rubric

The user prompt provides the scoring rubric and task instructions:

```
Your task is to evaluate each candidate
    response for each case.

Scoring Dimensions (1-5, integer):

1. Coherence
   - 1 = Unrelated, contradictory, or repeats
     context
   - 3 = Generally relevant but weakly grounded
   - 5 = Precisely grounded and context-aware

2. Empathy
   - 1 = Ignores or misreads emotion
   - 3 = Acknowledges emotion but generic
   - 5 = Emotionally precise without
     exaggeration

3. Informativeness
   - 1 = No new information or pure paraphrase
   - 3 = Minor added value
   - 5 = Substantive, relevant insight

4. Continuity
   - 1 = Dead-end response
   - 3 = Neutral continuation
   - 5 = Opens a natural conversational path

Calibration Rules:
- A score of 5 is rare and reserved for
    exceptional responses.
- Most reasonable responses should fall in the
    2-4 range.
- Responses that repeat the context must
    receive:
  coherence <= 2, Informativeness <= 2.

[Input Data Placeholder]
```

These prompts ensure consistent, calibrated evaluation across all responses, preventing score inflation and maintaining strict quality standards.