Inference-Time Alignment via Hypothesis Reweighting

Anonymous authors
Paper under double-blind review

Abstract

Chat assistants must handle diverse and often conflicting user preferences, requiring adaptability to various user needs. We propose Hypothesis Reweighting (HyRe), a method that enables real-time personalization by reweighting ensemble members based on just 1-5 labeled examples from the target user or domain. Our method builds on the key empirical observation that optimally weighting ensemble members substantially outperforms uniform averaging under distribution shift, providing a powerful inductive bias for personalization. HyRe trains a single network with multiple prediction heads that capture different valid interpretations of preference data, then performs a simple Bayesian update to upweight heads that best match the target user's preferences. This requires only a single forward pass with negligible (<1%) computational overhead, making it practical for inference-time alignment. We empirically validate HyRe in several target evaluation distributions. With as few as five preference pairs from each target distribution, adaptation via HyRe surpasses state-of-the-art reward models on RewardBench at both the 2B and 8B parameter scales, and improves reward model accuracy by 20% across 32 diverse personalization tasks.

1 Introduction

Task specification—describing precisely what a machine learning model should do—is inherently iterative and fundamentally incomplete under any finite set of instructions or training examples. As models grow more powerful and are applied to increasingly complex and nuanced tasks, this problem arises in many forms, from spurious correlations in the data to conflicting user preferences. Consider a chatbot trained via Reinforcement Learning from Human Feedback (RLHF) [61] on a broad distribution of user preferences. Such models often perform adequately in aggregate but systemtically fail to address specific users' needs, since different individuals have distinct, sometimes contradictory, notions of desirable responses. Meeting these user-specific requirements necessitates rapid model adaptation with minimal supervision.

However, existing adaptation strategies face practical constraints: fine-tuning requires hundreds of gradient steps and risks catastrophic forgetting [27, 28, 47, 71], while in-context learning methods scale poorly with context length and lack sample efficiency in few-shot regimes [21, 36, 74]. For applications requiring sub-second per-user adaptation—such as personalized assistants serving millions of concurrent users—these constraints become critical bottlenecks. This motivates inference-time methods that can adapt without gradient updates or large context windows.

To efficiently resolve ambiguity at test time, we draw on recent progress in efficient ensemble architectures [49]. These methods let a single backbone network represent a broad range of plausible functions at a small overhead, capturing the different ways the model can interpret the training set. While prior work focuses largely on using ensembles for uncertainty estimation, we propose using them to disambiguate tasks in real time: by quickly assessing which members of the ensemble best match a new distribution, we can "pick the right interpretation" for that scenario.

We introduce Hypothesis Reweighting (HyRe), a two-step approach that scales to large models. First, we train an ensemble of function heads on top of a shared backbone, ensuring each head individually fits the training data. Next, at inference time, we gather a few labeled examples from the target distribution—either proactively queried or provided in advance—and measure each head's performance. We then reweight the ensemble using a generalized Bayesian update that favors the heads performing best on the adaptation set.

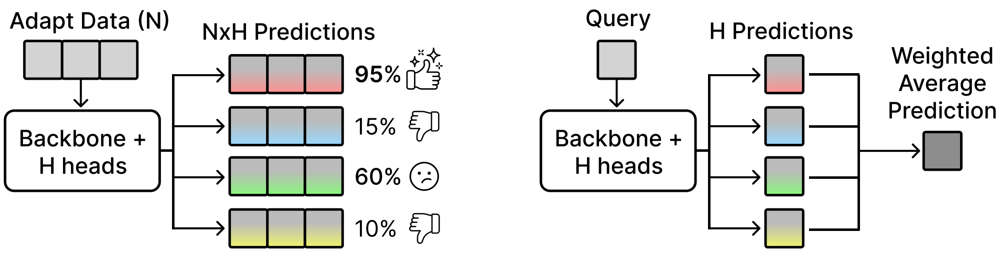


Figure 1: **Overview of HyRe**. We train multiple prediction heads on a shared backbone. (Left) At inference time, we evaluate each head on a small labeled adaptation set drawn from the target distribution. (Right) We reweight the heads according to the sum of their accuracies on the adaptation set, and use the weighted ensemble to make predictions on new inputs.

Crucially, this update supports non-differentiable metrics like 0-1 error, and it requires only a single forward pass over the adaptation set, making it far more efficient than conventional fine-tuning.

Our key contributions are as follows. We introduce HyRe, a simple and efficient method for inference-time adaptation that reweights ensemble members based on a few labeled examples from the target distribution. We demonstrate that uniform ensemble averaging often fails under distribution shift, motivating our adaptive reweighting approach. We achieve state-of-the-art results on RewardBench at both 2B and 8B scales with just 1-5 adaptation examples, outperforming much larger models. Finally, we show the method scales to large models with negligible overhead ($< 1\%$) and improves performance across 32 diverse preference learning tasks.

2 Preliminaries

Problem setup. We consider a general supervised learning setting that includes classification, preference learning, and regression tasks. Let $\mathcal{X}$ represent the input space and $\mathcal{Y}$ the output space, with training distribution $\mathrm{P}_{\text{train}}$ and evaluation distribution P_{eval} defined over $\mathcal{X} \times \mathcal{Y}$. The training dataset $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$ consists of N examples drawn from $\mathrm{P}_{\text{train}}$. We explore few-shot adaptation settings such as chatbot personalization, where a small adaptation set $\mathcal{D}_{\text{adapt}} \sim \mathrm{P}_{\text{eval}}$ only partially informs model performance under P_{eval}. The adaptation set $\mathcal{D}_{\text{adapt}}$ can be labeled in advance or actively queried, and is much smaller than the training set ($|\mathcal{D}_{\text{adapt}}| \ll |\mathcal{D}_{\text{train}}|$). For instance, in our main experiment, $|\mathcal{D}_{\text{adapt}}| = 16$ compared to $|\mathcal{D}_{\text{train}}| > 300,000$, with adaptation occurring near-instantly after a single forward pass through the network.

Ensemble architectures. We train an ensemble of K models $f_1, \dots, f_K$ on the training data $\mathcal{D}_{\text{train}}$. We consider parameterizations of the ensemble that aim to represent a distribution over functions by training multiple models on the same dataset $\mathcal{D}_{\text{train}}$, ensuring diversity without computational overhead beyond training a single model. To achieve this, we employ *prior networks* [49]: fixed, randomly initialized models whose outputs are added to each ensemble member's output. This mechanism preserves diversity among ensemble members during training, even as individual models converge. We consider two computationally efficient ensemble architectures:

1. **Shared-Base Ensemble**: A single neural network that parameterizes both the prior and ensemble components by sharing a common base.

2. **Epinet**: A base network augmented by a small auxiliary network that introduces diversity via a learned index.

We train all ensemble members jointly by minimizing $\sum_{k=1}^{K} \mathcal{L}(f_k, \mathcal{D}_{\text{train}})$ using SGD. Note that this simple training procedure does not explicitly encourage diversity and may lead to ensemble collapse in some settings. However, we empirically observe that sufficient diversity emerges for personalization tasks, likely due to random initialization and the high-dimensional parameter space of large models. These architectures have negligible overhead—in our reward model experiments, 100 ensemble heads add only 550K parameters (0.03%) to the 2B-parameter Gemma backbone. Please refer to Section D for architectural details.

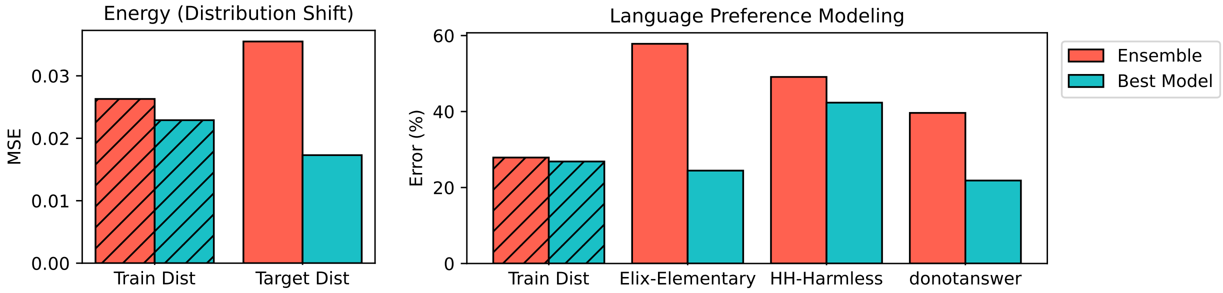


Figure 2: Motivating observation: the ensemble average often performs worse than a single well-chosen member. This tendency is particularly pronounced further away from the training distribution. **HyRe goes further than selecting the best head—it finds a continuous weighting of all heads**.

3 Inference-Time Ensemble Recalibration

In this section, we motivate and describe Hypothesis Reweighting (HyRe), a simple and computationally efficient method for few-shot adaptation to new tasks. HyRe dynamically adjusts the weights assigned to different ensemble members at test time based on a few labeled samples from the new task. HyRe leverages the ensemble's diversity—each member representing a different function that fits the training data—to efficiently adapt without retraining any model parameters.

3.1 Uniform Ensemble Averaging is Often Suboptimal

A long-held principle in machine learning is that a uniformly weighted ensemble of independently trained models tends to outperform individual models, largely because averaging can smooth out errors and mitigate individual biases [15, 24, 38, 39, 51]. By exploiting diversity among ensemble members, the uniform average often reduces variance and improves predictive performance in settings where the training data sufficiently specifies the task at hand.

However, settings with substantial distribution shifts or personalized requirements can render the uniform ensemble average overly coarse. When ensemble members represent different valid interpretations of the training data, averaging dilutes these meaningful distinctions. Recent work [42, 67] shows that in such cases, selecting the best single model can outperform naive averaging, as individual models may better align with specific test conditions.

While prior studies have shown this phenomenon in controlled or synthetic tasks, we verify in Figure 2 that it persists in large-scale, real-world settings involving distribution shifts and personalization. Uniform ensemble predictions indeed fall short of even the accuracy achieved by a suitably chosen single model.

This observation motivates our approach: *dynamically adjust ensemble weights based on alignment with the target task.* Using a small labeled set from the target distribution, we upweight the prediction heads that best match target data. We detail this reweighting procedure in Section 3.2 and validate its effectiveness on real-world tasks in Section 6.

Key insight: Uniform averaging fails under distribution shift

In distribution shift settings, the ensemble average often performs worse than the best single member. HyRe exploits this by finding optimal weightings that select the right interpretation for each context.

3.2 HyRe: Fast Inference-Time Ensemble Reweighting

Given an ensemble of K models $f_1, \ldots, f_K$, we aim to dynamically update their weights based on adaptation data. As a practical inference-time assumption in settings where we cannot further train neural networks, we can think of the "best" model as being one of the K ensemble particles that performs best on the evaluation distribution. Starting with uniform weights $w_k = \frac{1}{K}$, we update them as new labeled data from P_{eval} becomes available.

Algorithm 1 HyRe (Inference Time)

```
Require: Ensemble members $f_{1..K}$, unlabeled dataset $x_{1..N}$, query budget $B$
Initialize weights $w \leftarrow [\frac{1}{K}, \dots, \frac{1}{K}]$, query set $Q \leftarrow \emptyset$
for $i \leftarrow 1$ to $B$ do
    (Optional) Query label $y_n$ for $\arg\max_n c(x_n)$ where $c(x_n)$ is ensemble disagreement    (Section B)
    Compute accuracy $\text{acc}_k = \sum_{n \in Q} \text{acc}(f_k, x_n, y_n)$ for each $k$
    Update ensemble weight $w_k \propto \exp(\text{acc}_k + p)$    (Section 3.2)
end for
Return final weighted ensemble function $f_w : x \mapsto \sum_{k=1}^{K} w_k f_k(x)$
```

The weighted ensemble prediction is $f_w(x) = \sum_{k=1}^{K} w_k f_k(x)$, where each $w_k \geq 0$ and $\sum_{k=1}^{K} w_k = 1$. We measure each member's performance using a loss function $l(f_k, x, y)$ and compute their cumulative loss on adaptation data $\mathcal{L}(f_k, \mathcal{D}_{\text{adapt}}) = \sum_{(x,y) \in \mathcal{D}_{\text{adapt}}} l(f_k, x, y)$. The weights are updated using a softmax on negative cumulative loss:

$$w_k = \frac{\exp(-\mathcal{L}(f_k, \mathcal{D}_{\text{adapt}}))}{\sum_{j=1}^{K} \exp(-\mathcal{L}(f_j, \mathcal{D}_{\text{adapt}}))}. \tag{1}$$

This update has a natural interpretation as generalized Bayesian inference (Section 3.3), where we treat each ensemble member as a hypothesis and update beliefs based on target-domain performance rather than likelihood. As the loss $l(f_k, x, y)$, we use 0-1 error for classification and mean squared error for regression, though HyRe supports any performance metric since the weight update remains valid for non-differentiable functions.

The complete adaptation procedure is summarized in Algorithm 1. Key practical considerations include:

- **Computational efficiency.** Since we use efficient ensemble architectures (Section 2), training and inference cost is comparable to that of a single network. Reweighting requires just one forward pass.
- **Coverage of $f_1, \dots, f_K$.** The target function need not lie exactly in the ensemble's linear span. Restricting solutions to the ensemble's convex hull provides a practical bias-variance tradeoff in low-data regimes.
- **Active selection of $\mathcal{D}_{\text{adapt}}$.** When unlabeled samples are available at test time, we can actively select which to label, improving reweighting efficiency (Section B).

3.3 Interpreting HyRe as Generalized Bayesian Inference

The weight update in (1) can be interpreted as a form of generalized Bayesian inference [7]. Given an initial belief state $\pi(w)$, the updated belief after observing $\mathcal{D}_{\text{adapt}}$ is:

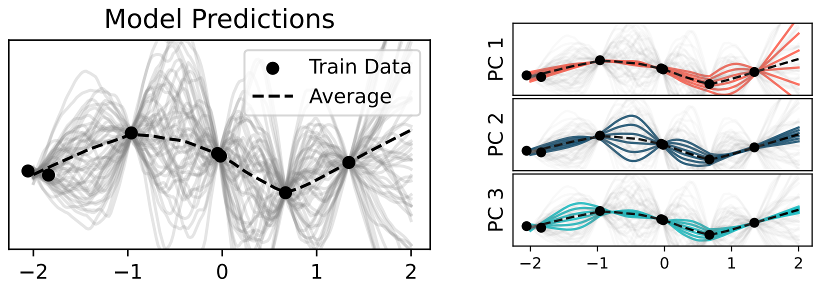
$$\pi(w | \mathcal{D}_{\text{adapt}}) \propto \exp\left(-\mathcal{L}(w, \mathcal{D}_{\text{adapt}})\right) \pi(w), \tag{2}$$

which generalizes classical Bayesian inference by allowing arbitrary loss functions. Standard Bayes is recovered when $l(w, x)$ is the negative log-likelihood.

Under mild conditions like i.i.d. sampling, these updates are consistent and coherent: they converge to the optimal weighting and yield identical posteriors whether applied incrementally or in batches. For classification tasks, using 0-1 loss instead of log-likelihood provides more stable updates by avoiding outlier dominance [30]. This makes HyRe particularly suitable for robust adaptation with non-differentiable metrics.

Takeaway: HyRe performs Bayesian inference given adaptation data.

HyRe performs a simple weight update based on target data accuracy (1), which has a natural interpretation as a generalized Bayesian inference procedure in the space of ensemble weights (2).



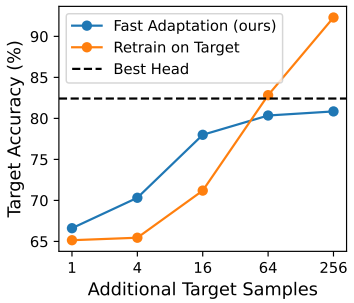


Figure 3: Principal component analysis of an ensemble of regression models. Left: Each gray line is the prediction of an ensemble member; the dashed line shows the ensemble mean. Right: The top three principal components of the ensemble's predictions reveal distinct axes of variation in predictive behavior. **Searching among ensemble weights like HyRe acts as a strong inductive bias** **nsis**

Figure 4: HyRe vs. fine-tuning with different amounts of adaptation data. Despite using only a single forward pass, **HyRe outperforms fine-tuning in low-data regimes**

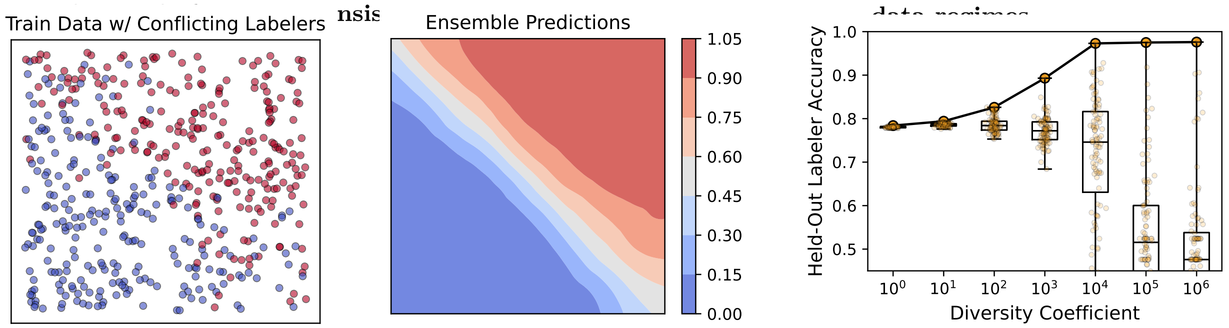


Figure 5: Ensemble behavior under label ambiguity. (Left) We simulate conflicting preferences between labelers on a synthetic dataset. (Center) Ensemble-averaged predictions approximate the consensus, smoothing over disagreements. (Right) We measure the maximum agreement between an ensemble and a held-out labeler: **increasing model diversity improves alignment with individual labelers**.

4 When is Ensemble Reweighting Effective, and Why?

This section explores *the conditions under which ensemble reweighting is effective* through three illustrative examples: analyzing ensemble diversity through PCA, examining decision boundaries in classification, and comparing adaptation strategies.

Ensemble diversity reflects task ambiguity. We visualize how an ensemble's diversity reflects the axes of task ambiguity. We consider a synthetic regression task where the training data is sampled from a Gaussian Process (GP) prior. For target inputs $x_1, \ldots, x_M$, each ensemble member f_k produces predictions $v_k \in \mathbb{R}^M$. We perform Principal Component Analysis (PCA) on the prediction matrix $V = (v_1, \ldots, v_K) \in \mathbb{R}^{M \times K}$ to yield components $u_1, \ldots, u_m \in \mathbb{R}^M$ that capture the main variations between ensemble members.

Using an ensemble of 100 models trained on 7 inputs and evaluated on 1000 test inputs, we visualize the first three principal components in Figure 3. Each component represents a distinct mode of variation while preserving smoothness and fit to training data. Like wavelets, these components are localized in input space and form a basis for approximating the ensemble. See Section F for further analysis of PCA applied to ensemble predictions.

Ensembles as diverse sharp decision boundaries. We build on an alternative interpretation of the Bradley-Terry model, where the model can be seen as representing a population of deterministic decision-makers. For items i and j with parameters $\theta_i, \theta_j \in \mathbb{R}$, the preference probability under the Bradley-Terry model is:

$$P(i \succ j) = \frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j}} = P\left(\theta_i + \epsilon_i > \theta_j + \epsilon_j\right), \tag{3}$$

where $\epsilon_i, \epsilon_j \sim \text{Gumbel}(0, 1)$. Rather than a single stochastic decision-maker, the model can be seen as representing a population of deterministic decision-makers. Each decision-maker is characterized by a pair (ϵ_i, ϵ_j), and makes sharp choices based on which among $\theta_i + \epsilon_i$ and $\theta_j + \epsilon_j$ is larger. The model's probabilistic behavior emerges from averaging across this population.

Method	**Energy**	**Kin8nm**	**CCPP**
MC Dropout	0.3033	0.6494	0.3761
Vanilla	0.1664	0.4514	0.2920
+ HyRe	0.1572 (-0.0092)	0.4498 (-0.0016)	0.2902 (-0.0018)
Epinet	0.1396	0.4823	0.3068
+ HyRe	0.1345 (-0.0051)	0.4814 (-0.0009)	0.3036 (-0.0032)
Shared-Base	0.1508	0.5316	0.2976
+ HyRe	0.1431 (-0.0077)	0.5314 (-0.0002)	0.2955 (-0.0021)

Table 1: RMSE (lower is better) on test data with distribution shifts across three UCI datasets. We compare the performance of various ensemble architectures with test-time adaptation using HyRe. We find that **for all three ensemble architectures, HyRe is consistently able to adapt to the distribution shift between training and test data**.

Model	**Helpful**	**Harmless**
Fine-Tune (Helpful)	73.03	32.59
Fine-Tune (Harmless)	32.06	73.30
Pretrained RM	68.01	52.16
Ensemble	66.34	50.90
+ HyRe (Helpful)	**68.44**	51.21
+ HyRe (Harmless)	64.24	**57.66**

Table 2: Helpful vs harmless tradeoff. To establish an upper bound on performance, we fine-tune the reward model on the helpful and harmless datasets separately. **Reweighting an ensemble model with HyRe allows us to flexibly trade off between the two desiderata**.

We hypothesize that diverse ensembles can learn such sharp decision boundaries from aggregate data across a population of annotators. To test this, we construct a synthetic preference learning task with conflicting labelers. We sample inputs (x_1, x_2) from $[0,1]^2$ and generate diverse linear decision boundaries $w_1 x_1 + w_2 x_2 > 0$, with $w_1, w_2 \sim N(0,1)$. As shown in Figure 5, our ensemble quickly adapts to new decision boundaries, outperforming single models. The average ensemble prediction matches the "average" decision-maker, while individual members capture distinct boundaries. In particular, higher diversity coefficients for the prior network yields sharper boundaries per ensemble member. In Section 6, we show this enables rapid personalization in real-world preference tasks.

HyRe outperforms fine-tuning in low-data regimes. We compare HyRe to model fine-tuning on a synthetic binary classification task. The training set contains inputs from $[0,1]^5$ labeled as 1 and inputs from $[-1,0]^5$ labeled as 0. The target distribution is uniform over $[-1,1]^5$ with a random linear decision boundary. Results in Figure 4 show that HyRe outperforms fine-tuning in the low-data regime, achieving high accuracy with few queries. Fine-tuning eventually surpasses reweighting with more data due to its higher capacity. This illustrates a bias-variance tradeoff: reweighting reduces variance by restricting solutions to the ensemble's span, providing an advantage with limited data. Additionally, HyRe requires only a single forward pass and negligible weight computation cost (1), making it especially suitable for large models and resource-constrained settings.

Key insight: Ensembles capture multiple valid interpretations

Ensembles naturally capture multiple valid interpretations of training data. Reweighting selects the right combination of interpretations for each new context without retraining.

5 Related Work

Ensembles. A long-standing theme in machine learning is using ensembles to improve predictive performance and uncertainty estimates when different members make independent mistakes [38, 39]. Our approach builds on efficient ensemble methods with shared backbones [49] and extends prior work on dynamic ensemble weighting [33, 56], though these typically focus on differentiable loss-based objectives during training. In contrast, we perform adaptive reweighting at inference time using minimal labeled data from the target distribution, supporting non-differentiable evaluation metrics for effective alignment. Recent work in multi-objective optimization [25, 45, 46] uses Chebyshev scalarization with exponential weighting for balancing training objectives; while our weighting scheme is similar, we apply it to test-time adaptation rather than training-time optimization.

Mixture-of-Experts. While our use of multiple prediction heads may superficially resemble Mixture-of-Experts (MoE) [16, 32, 43, 58], the two approaches serve fundamentally different purposes. MoE models route inputs to specialized sub-networks via sparse activation, performing conditional computation within a single model for efficiency. In contrast, HyRe evaluates all heads on every input to maintain a posterior distribution over complete hypotheses, enabling uncertainty quantification via Bayesian model averaging rather than computational savings.

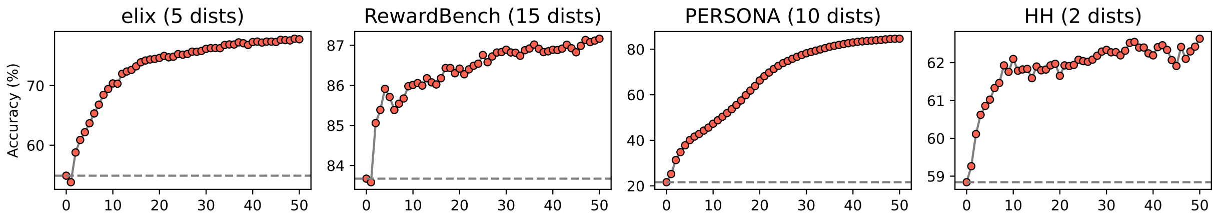


Figure 6: Average reward model accuracy as a function of adaptation set size. The dashed line shows the best available static 2B reward model for each dataset group. **HyRe consistently outperforms the state-of-the-art reward model with as few as 1-5 examples per distribution**.

Task underspecification and scalable alignment. In many machine learning tasks, the training data fails to fully define desired model behavior [14, 22]. This challenge intensifies under limited data or distribution shifts, where multiple hypotheses remain consistent with observations. Reinforcement learning faces similar issues: reward specification is difficult in open-ended environments, and optimizing misspecified objectives can lead to unintended behaviors [20, 52, 62, 76]. Instead of fully defining a task upfront, one can collect human demonstrations or pairwise preferences, framing task specification as a cooperative game between agents and humans [23]. Reinforcement Learning from Human Feedback (RLHF) operationalizes this idea by using user preferences to guide post-training [11, 50, 54, 70], with some using ensembles [2, 12, 75]. Recent work on pluralistic alignment [63] uses explicit domain labels or per-user data to improve personalization [6, 10, 31, 44, 53]. However, these methods require explicit domain labels or per-user data. Specialized inference-time alignment approaches [17, 29, 34, 48] focus on token-level steering or reranking but assume a single fixed reward and cannot reconcile conflicting user preferences at test time. HyRe demonstrates that this additional information is not necessary during training: a diverse ensemble trained on aggregate data can capture ambiguity, which we can use to directly adapt to new users. Our experiments show this insight generalizes across several problem settings.

6 Experiments

We now empirically validate HyRe. We focus on three key questions: (1) Can HyRe effectively handle mild covariate shift? (2) Does HyRe scale to large models? (3) How robust and computationally efficient is HyRe? We describe the detailed setup for each experiment in the appendix.

6.1 Regression Data with Mild Covariate Shift

We evaluate HyRe on three UCI regression datasets [35]—Energy Efficiency, Kin8nm, and CCPP—using the protocol of Sharma et al. [57]: the top and bottom 5% of the data (sorted by mean input features) form an OOD target set, while the central 90% is split into train and validation sets. All methods employ 100 two-layer MLPs with 50 units each. As baselines, we consider a vanilla ensemble of independently trained models and MC Dropout [18]. We report the best-performing MC Dropout results across all architectures. Results in Table 1 demonstrate that uniform ensembles perform strongly in these OOD generalization settings and that HyRe consistently improves over the uniform ensemble.

6.2 Scalable Personalization of Preference Models

Experimental setup. We evaluate personalization using four sets of human preference benchmarks: Elix [60], RewardBench [40], PERSONA [9], and Anthropic HH [5]. Together, these benchmarks contain 32 datasets, each encoding a different aspect of human preferences. To train HyRe on preference data, we attach Shared-Base ensemble heads to a pretrained 2B reward model and fine-tune it on the UltraFeedback [13] dataset, a standard dataset for reward model training. We use two public finetune checkpoints of Gemma 2B models, which achieve state-of-the-art performance on RewardBench at the 2B parameter scale, even outperforming GPT-4o [1]. Refer to Section C for our detailed setup.

We first evaluate the effectiveness of HyRe in adapting our reward model ensemble to new distributions at test time, comparing its performance to that of the original reward model. As shown in Figure 6, a simple

Model	Type	Overall	Chat	Chat Hard	Safety	Reasoning
Tulu-2-DPO-70B	DPO	79.1	97.5	60.5	84.5	74.1
Claude-3 Sonnet (June 2024)	Gen	84.2	96.4	74.0	81.6	84.7
GPT-4 (Aug 2024)	Gen	86.7	96.1	76.1	88.1	86.6
Gemini-1.5-Pro-0924	Gen	86.8	94.1	77.0	85.8	90.2
INF-ORM-Llama3.1-70B	Seq	95.1	96.6	91.0	93.6	99.1
GRM-Gemma2-2B	Seq	88.4	93.0	**77.2**	92.2	91.2
+ Ours (uniform)	Seq	87.1	96.4	73.1	87.4	89.8
+ Ours (N=1)	Seq + HyRe	86.5	92.4	71.5	85.1	92.5
+ Ours (N=5)	Seq + HyRe	88.5	95.0	72.5	90.3	93.1
+ Ours (N=10)	Seq + HyRe	**89.7**	**96.4**	74.7	**92.4**	**93.5**
+ Ours (best weight oracle)*	Seq + Oracle	93.1	98.3	83.4	96.7	94.9
Skywork–Llama-3.1-8B	Seq	94.0	94.7	88.6	92.7	96.7
+ Ours (uniform)	Seq	94.0	95.0	87.2	93.0	96.8
+ Ours (N=1)	Seq + HyRe	94.3	95.2	87.8	93.0	97.5
+ Ours (N=5)	Seq + HyRe	94.7	95.5	88.6	93.2	97.8
+ Ours (N=10)	Seq + HyRe	**95.0**	**95.9**	**89.3**	**93.5**	**97.9**
+ Ours (best weight oracle)*	Seq + Oracle	97.2	99.2	93.0	96.5	98.8

∗ *Oracle* methods show an upper bound on performance, using the test set.

Table 3: Accuracy across tasks in RewardBench. We report overall performance and breakdowns by task category for all models. **HyRe improves upon the state-of-the-art models at the 2B and 8B parameter scales with as few as 1-5 labeled samples per distribution**.

Dataset	N=0	N=1	N=5	N=10	N=20	N=40	N=80
GPT-4o-mini N-Shot Prompting							
donotanswer	44.4	50.3	60.2	64.7	68.7	66.4	67.0
refusals	79.4	82.7	80.5	82.2	82.1	78.0	82.1
Llama-3.1-8B N-Shot Prompting							
donotanswer	46.6	52.8	59.6	63.4	41.2	62.6	*
refusals	61.6	82.4	80.0	72.2	44.4	79.0	*
Llama-3.1-8B + HyRe							
donotanswer	58.6	60.8	69.1	**71.3**			
refusals	88.9	90.0	94.0	**95.2**			

Table 4: Comparison with few-shot prompting on two datasets from RewardBench. (*) exceeds Together AI API token limit. We see a degradation in performance for both GPT-4o-mini and Llama-3.1-8B as we increase the number of examples, whereas HyRe consistently outperforms both across all sample sizes. **HyRe provides reliable test-time alignment, unlike few-shot prompting, which can degrade with too much context**.

	donotanswer	xstest-sr	refusals
HyRe w/ Cross-Entropy			
N=0	58.60 ± 4.93	82.80 ± 2.43	88.90 ± 3.25
N=1	62.53 ± 4.00	85.78 ± 3.06	92.49 ± 3.28
N=5	62.57 ± 1.88	87.38 ± 1.26	93.17 ± 2.19
N=10	62.25 ± 1.71	87.51 ± 1.19	93.21 ± 1.77
HyRe w/ Accuracy			
N=0	58.60 ± 4.93	82.80 ± 2.43	88.90 ± 3.25
N=1	60.81 ± 5.29	85.80 ± 2.45	90.00 ± 3.63
N=5	69.12 ± 5.81	89.28 ± 2.86	94.00 ± 2.45
N=10	**71.32 ± 6.33**	**90.32 ± 3.20**	**95.20 ± 2.32**
Oracle	76.54 ± 2.35	90.32 ± 1.91	99.50 ± 0.87

Table 5: Cross-entropy ablation experiment. We report average and std of accuracy (%) with varying numbers of adaptation examples (N) on three datasets. **Using accuracy as the adaptation objective for HyRe significantly improves post-adaptation performance**.

uniform ensemble initially underperforms the original model, indicating that naive ensembling alone cannot ensure broad generalization. Nevertheless, HyRe quickly surpasses the baseline with just a few labeled examples per distribution. We show detailed dataset-level results in the appendix (Figure 9).

We compare HyRe against state-of-the-art reward models on the RewardBench leaderboard at both the 2B and 8B parameter scales. As shown in Table 11, HyRe—with only 1-5 labeled examples per distribution—exceeds the performance of many much larger reward models. We note that these reward models outperform strong generative reward models including Claude 3.5 Sonnet, GPT-4, and Gemini-1.5-Pro [1, 3, 66]. This indicates that inference-time alignment can be a powerful alternative to naively scaling up reward models.

Method	Win Rate
Random	50.0%
Uniform	61.4%
HyRe (n=1)	60.0%
HyRe (n=2)	61.7%
HyRe (n=4)	62.4%
HyRe (n=8)	63.1%
HyRe (n=16)	63.5%

Table 6: Best-of-N win rates on Summarize from Feedback. **HyRe improves output quality with more examples.**

Ratio	`math-prm`	`xstest-sr`
0.0 : 1.0	72.57%	88.33%
0.1 : 0.9	98.94%	86.64%
0.2 : 0.8	96.73%	88.15%
0.5 : 0.5	98.52%	87.22%
0.8 : 0.2	99.38%	86.66%
0.9 : 0.1	99.52%	85.86%
1.0 : 0.0	99.72%	84.18%

Table 7: Accuracy with mixed adaptation data. **HyRe recovers 97% accuracy despite extreme imbalance.**

Method	Acc (%)
Single Model	59.03
Entropy Weighted	68.38
Logit Ensemble [33]	83.44
Majority Vote	83.71
GEM [56, N=40]	84.49
GEM [56, Oracle†]	89.51
HyRe (N=1)	83.88
HyRe (N=5)	85.73
HyRe (N=10)	86.26
HyRe (N=20)	87.11
HyRe (N=40)	**87.74**

Table 8: Comparison of ensemble aggregation methods on RewardBench. **HyRe outperforms all methods with only 5 examples.**

Takeaway: HyRe outperforms state-of-the-art reward models.

With just 1-5 labeled datapoints per distribution, HyRe outperforms state-of-the-art reward models on RewardBench at both the 2B and 8B parameter scales.

Beyond standard reward model accuracy, we validate that HyRe improves actual text generation quality using best-of-N sampling on the OpenAI Summarize from Feedback dataset [64], where multiple model-generated summaries are ranked by human Likert scores. As shown in Table 6, HyRe consistently outperforms the uniform ensemble baseline and scales with adaptation examples, translating reward model improvements into measurable downstream quality gains (McNemar's test p-values 10^{-16} to 10^{-3} across 4000+ comparisons).

6.3 Comparison with Alternative Adaptation Methods

Fine-tuning on target data. We compare HyRe against models fine-tuned on the helpful-base and harmless-base training sets in the Anthropic-HH dataset. Results in Table 2 indicate that while targeted fine-tuning models achieve higher performance in their respective target metrics, they significantly reduce performance in the other. In contrast, our HyRe-adapted ensemble not only increases performance across each data distribution but also retains or slightly improves performance in the other split. We emphasize that we show fine-tuning performance only as a point of comparison; **fine-tuning a model for a target distribution is usually too computationally expensive to be done at inference time, and is thus not a practical solution for inference-time alignment**.

Few-shot prompting. We compare HyRe with few-shot prompting using GPT-4o-mini on two datasets from RewardBench. As shown in Table 4, HyRe consistently outperforms GPT-4o-mini across all sample sizes. Even in the zero-shot setting, specialized reward models like HyRe achieve higher performance than general-purpose language models like GPT-4o-mini. Notably, we observe that too many few-shot examples can actually harm performance, as seen with GPT-4o-mini's performance drop after N=10 for donotanswer and N=20 for refusals. These results demonstrate that specialized reward models with inference-time adaptation can more efficiently leverage few-shot examples than general-purpose language models.

6.4 Analysis and Ablation Studies

Computational overhead. HyRe uses a single pre-trained backbone with K small prediction heads, ensuring minimal parameter overhead. In our reward model experiments, 100 ensemble heads (550K parameters) add only 0.03% to the Gemma-2B backbone. At inference time, reweighting requires a single forward pass through the backbone and heads, with negligible weight computation. The total cost increase in time and memory is less than 1%, making the approach practical for deployment in production systems where fine-tuning would be prohibitively expensive.

Ablation on reweighting criteria. We compare accuracy-based reweighting against binary cross-entropy loss. Cross-entropy significantly degrades performance across three RewardBench datasets (Table 5) because it is unbounded and sensitive to outliers, causing rapid overfitting to individual heads. This validates our use of generalized Bayesian inference with non-differentiable metrics rather than likelihood-based updates.

Robustness to skewed adaptation data. We test HyRe on non-i.i.d. scenarios by mixing two RewardBench datasets (math-prm and xstest-should-respond) at different ratios. We train on these different mixtures, then evaluate the resulting ensemble on each distribution separately. Even with highly skewed mixtures (e.g., only 10% from dataset A), HyRe recovers 97% of the accuracy gains compared to adapting exclusively on dataset A (Table 7). This robustness to noisy or mixed adaptation signals is crucial for real-world deployment, where users may provide inconsistent or multi-domain feedback.

Comparison with alternative reweighting methods. We compare HyRe against existing ensemble weighting approaches: uniform weighting [33], confidence weighting, majority voting, and convex optimization [56]. HyRe consistently outperforms all prior methods with just 1-5 examples per distribution (Table 8), demonstrating that our update is more sample-efficient than existing ad-hoc weighting heuristics.

Practical advantages of HyRe

HyRe adds only 0.03% parameters (100 heads on 2B model) and $< 1\%$ computational overhead. Reweighting requires just one forward pass vs. hundreds of gradient steps for fine-tuning, making it practical for real-time adaptation.

7 Discussion

Our results demonstrate how efficient ensemble architectures can offer a practical path to inference-time alignment in large models. The proposed method, HyRe, is well-suited for settings where fine-tuning is impractical, such as regulatory-frozen models in medical AI that require recertification for any parameter change, real-time personalization for millions of users under strict latency constraints, and live content moderation where community norms drift faster than batch retraining cycles. By attaching lightweight ensemble “heads” to a shared backbone, we can capture multiple plausible interpretations of the training distribution at negligible extra cost. Then, through a simple reweighting step that leverages just a handful of target-domain examples, the ensemble can effectively pick out the functions that align best with a new task. A natural next step is to close the loop by pairing our approach with a parameterization of the reward model that allows for direct behavior adjustments [54].

Our method currently relies on a small batch of labeled examples from the target distribution, and does not address single-sample or online streaming adaptation. Furthermore, while relying on minimal data, our reweighting still assumes that the ensemble’s functional diversity covers the new domain’s core behaviors. Extending our framework to dynamically expand or augment the ensemble as new tasks emerge is an exciting direction. Nevertheless, our results demonstrate that lightweight ensembles with inference-time reweighting offer a promising and practical approach for aligning large models at inference time.

Scope and applicability. HyRe demonstrates strong improvements on personalization tasks, where ensemble diversity naturally captures different valid interpretations of the same underlying task. On natural distribution shifts such as those in WILDS [37], gains are more modest (see Table 12), suggesting the method is most effective when the target domain lies within the functional diversity learned by the ensemble. This highlights a natural trade-off: training ensembles to cover diverse interpretations of preference data makes them well-suited for personalization, while more severe covariate shifts may require different adaptation strategies.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. [page 7, 8]

[2] Ahmed M Ahmed, Rafael Rafailov, Stepan Sharkov, Xuechen Li, and Sanmi Koyejo. Scalable ensembling for mitigating reward overoptimisation. *arXiv preprint arXiv:2406.01013*, 2024. [page 7]

[3] Anthropic. Claude 3.5 sonnet. Accessed via Claude.ai, API, and cloud platforms, 2024. URL https://www.anthropic.com. Enhanced reasoning, state-of-the-art coding skills, computer use, and 200K context window. Available on Anthropic API, Amazon Bedrock, and Google Cloud's Vertex AI. [page 8]

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. [page 18]

[5] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. [page 7, 20]

[6] André Barreto, Vincent Dumoulin, Yiran Mao, Nicolas Perez-Nieves, Bobak Shahriari, Yann Dauphin, Doina Precup, and Hugo Larochelle. Capturing individual human preferences with reward features. *arXiv preprint arXiv:2503.17338*, 2025. [page 7]

[7] P. G. Bissiri, C. C. Holmes, and S. G. Walker. A General Framework for Updating Belief Distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1103–1130, 02 2016. ISSN 1369-7412. doi: 10.1111/rssb.12158. [page 4]

[8] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018. [page 17]

[9] Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. Persona: A reproducible testbed for pluralistic alignment. *arXiv preprint arXiv:2407.17387*, 2024. [page 7, 20, 24]

[10] Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Korlakai Vinayak. Modeling the plurality of human preferences via ideal points. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*, 2024. URL https://openreview.net/forum?id=YkdoxVxy6a. [page 7]

[11] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017. [page 7]

[12] Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*, 2023. [page 7]

[13] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023. [page 7]

[14] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. [page 7]

[15] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pp. 1–15. Springer, 2000. [page 3]

[16] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. [page 6]

[17] Yu Fei, Yasaman Razeghi, and Sameer Singh. Nudging: Inference-time alignment of llms via guided decoding. *arXiv preprint arXiv:2410.09300*, 2024. [page 7]

[18] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016. [page 7]

[19] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pp. 1183–1192. PMLR, 2017. [page 17]

[20] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023. [page 7]

[21] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020. [page 1]

[22] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. [page 7]

[23] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016. [page 7]

[24] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990. [page 3]

[25] Yifei He, Shiji Zhou, Guojun Zhang, Hyokun Yun, Yi Xu, Belinda Zeng, Trishul Chilimbi, and Han Zhao. Robust multi-task learning with excess risks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 18151–18178. PMLR, 2024. [page 6]

[26] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011. [page 17, 23, 24]

[27] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019. [page 1]

[28] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. [page 1]

[29] James Y Huang, Shuyang Shawn Gu, Le Yu, Yuexin Wang, Jiaqi Ma, Hongkun Zhou, Jiayi Wang, Yutao Zhu, Siqi Jian, Yan Xie, et al. Deal: Decoding-time alignment for large language models. *arXiv preprint arXiv:2402.06147*, 2024. [page 7]

[30] Pavel Izmailov, Patrick Nicholson, Sanae Lotfi, and Andrew G Wilson. Dangers of bayesian model averaging under covariate shift. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 3309–3322. Curran Associates, Inc., 2021. [page 4]

[31] 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. *arXiv preprint arXiv:2310.11564*, 2023. [page 7]

[32] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. [page 6]

[33] D. Jimenez. Dynamically weighted ensemble neural networks for classification. In *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227)*, volume 1, pp. 753–756 vol.1, 1998. doi: 10.1109/IJCNN.1998.682375. [page 6, 9, 10]

[34] Yuu Jinnai, Joon Sung Yoo, Lam M Vu, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Regularized best-of-n sampling with minimum bayes risk objective for language model alignment. *arXiv preprint arXiv:2404.01054*, 2024. [page 7]

[35] Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. Uci machine learning repository. URL https://archive.ics.uci.edu. Accessed October 2024. [page 7]

[36] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023. [page 1]

[37] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021. [page 10, 16, 17]

[38] Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In G. Tesauro, D. Touretzky, and T. Leen (eds.), *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1994. [page 3, 6]

[39] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. [page 3, 6, 23]

[40] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling, 2024. [page 7, 20]

[41] Michael Laskin, Denis Yarats, Hao Liu, Kimin Lee, Albert Zhan, Kevin Lu, Catherine Cang, Lerrel Pinto, and Pieter Abbeel. Urlb: Unsupervised reinforcement learning benchmark. *arXiv preprint arXiv:2110.15191*, 2021. [page 17]

[42] Yoonho Lee, Huaxiu Yao, and Chelsea Finn. Diversify and disambiguate: Learning from underspecified data. *International Conference on Learning Representations*, 2023. [page 3, 22]

[43] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020. [page 6]

[44] Xinyu Li, Zachary C Lipton, and Liu Leqi. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*, 2024. [page 7]

[45] Xi Lin, Xiaoyuan Zhang, Zhiyuan Yang, Fei Liu, Zhenkun Wang, and Qingfu Zhang. Smooth tchebycheff scalarization for multi-objective optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 30479–30509. PMLR, 2024. [page 6]

[46] Meitong Liu, Xiaoyuan Zhang, Chulin Xie, Kate Donahue, and Han Zhao. Online mirror descent for tchebycheff scalarization in multi-objective optimization. *arXiv preprint arXiv:2410.21764*, 2024. [page 6]

[47] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024. [page 1]

[48] Zhixuan Liu, Zhanhui Yang, Huizhuo Qiu, Yichen Wang, Chengzhengxu Zhou, Zhichen Yu, I-Chun Chern, Joshua Susskind, and Navdeep Jaitly. Inference-time language model alignment via integrated value guidance. *arXiv preprint arXiv:2409.17819*, 2024. [page 7]

[49] Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Epistemic neural networks. *Advances in Neural Information Processing Systems*, 36, 2023. [page 1, 2, 6, 22]

[50] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. [page 7]

[51] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. [page 3]

[52] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv preprint arXiv:2201.03544*, 2022. [page 7]

[53] 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. [page 7]

[54] 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, 2024. [page 7, 10]

[55] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. [page 18]

[56] Mohsen Shahhosseini, Guiping Hu, and Hieu Pham. Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. *Machine Learning with Applications*, 7:100251, 2022. [page 6, 9, 10]

[57] Mrinank Sharma, Sebastian Farquhar, Eric Nalisnick, and Tom Rainforth. Do bayesian neural networks need to be fully stochastic?, 2023. [page 7]

[58] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. [page 6]

[59] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021. [page 18]

[60] Anikait Singh, Sheryl Hsu, Kyle Hsu, Eric Mitchell, Stefano Ermon, Tatsunori Hashimoto, Archit Sharma, and Chelsea Finn. Fspo: Few-shot preference optimization of synthetic preference data in llms elicits effective personalization to real users. *arXiv preprint arXiv:2502.19312*, 2025. [page 7, 19]

[61] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in rlhf. *arXiv preprint arXiv:2312.08358*, 2023. [page 1]

[62] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022. [page 7]

[63] Taylor Sorensen, Jared Sarkar, Sarah Kreps, and Christopher Lopatka. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024. [page 7]

[64] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Advances in neural information processing systems*, volume 33, pp. 3008–3021, 2020. [page 9]

[65] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. *Domain adaptation in computer vision applications*, pp. 153–171, 2017. [page 18]

[66] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. [page 8]

[67] Damien Teney, Ehsan Abbasnejad, Simon Lucey, and Anton van den Hengel. Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16761–16772, June 2022. [page 3, 22]

[68] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622, 1999. [page 24]

[69] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020. [page 19]

[70] Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017. [page 7]

[71] Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Reft: Representation finetuning for language models. *arXiv preprint arXiv:2404.03592*, 2024. [page 1]

[72] Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. Regularizing hidden states enables learning generalizable reward model for llms. *arXiv preprint arXiv:2406.10216*, 2024. [page 18]

[73] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pp. 25407–25437. PMLR, 2022. [page 16, 18]

[74] Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*, 2024. [page 1]

[75] Shun Zhang, Zhenfang Chen, Sunli Chen, Yikang Shen, Zhiqing Sun, and Chuang Gan. Improving reinforcement learning from human feedback with efficient reward model ensemble. *arXiv preprint arXiv:2401.16635*, 2024. [page 7]

[76] Simon Zhuang and Dylan Hadfield-Menell. Consequences of misaligned ai. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15763–15773. Curran Associates, Inc., 2020. [page 7]

Method	**N=0**	**N=1**	**N=5**	**N=10**	**N=20**	**N=40**
HyRe + Random	84.40	85.33	86.97	87.34	88.01	88.83
HyRe + Entropy	84.40	84.25	86.73	87.54	88.60	89.76
HyRe + BALD	84.40	84.28	87.13	87.78	88.60	88.99

Table 9: Accuracies on RewardBench with different datapoint selection strategies. While active sampling methods perform slightly better, **even random sampling consistently improves performance with the HyRe reweighting process**.

Algorithm	DL	Test Acc
IRM	O	64.2 (8.1)
CORAL	O	59.5 (7.7)
Group DRO	O	68.4 (7.3)
Fish	O	74.7 (7.1)
LISA	O	77.1 (6.9)
ERM	X	70.3 (6.4)
Evading	X	73.6 (3.7)
Ensemble	X	71.5 (3.4)
Ensemble + HyRe	X	**75.2** (5.3)

Table 10: Test set accuracy on Camelyon17. HyRe achieves competitive performance without using domain labels (DL).

A Additional Experiments

Effect of sampling strategy. In Table 9, we compare the performance of different active learning criteria for selecting adaptation data points. We consider random sampling, BALD, and entropy, measuring their performance over 0 to 40 target examples. Across the acquisition of 40 examples, active learning methods (BALD and entropy) demonstrated slightly better performance compared to random sampling. Even random sampling consistently improves performance, indicating that HyRe can be used with data collected before inference without sacrificing performance.

WILDS experiments. We evaluate a trained Shared-Base ensemble, both with and without HyRe on the WILDS-Camelyon17 dataset [37], comparing against several representative methods for OOD generalization from the official WILDS benchmark. As shown in Table 10, test-time adaptation with HyRe consistently outperforms other methods that do not use domain labels and remains competitive with LISA [73], a strong method that leverages domain labels for targeted data augmentation. We also test Shared-Base ensembles on four additional WILDS datasets (CivilComments, Amazon, FMoW, iWildCam), but did not observe further improvements from ensemble reweighting via HyRe, as detailed in Table 12. Nonetheless, training a diverse ensemble consistently improved OOD generalization in these datasets. We attribute the limited benefit of ensemble reweighting in these cases to some natural distribution shifts behaving similarly to in-distribution data in terms of task underspecification. For further discussion on the conditions that can make a single model outperform the ensemble, see Section 4.

We further compare the performance of HyRe with few-shot fine-tuning with the same amount of adaptation data. We evaluate both HyRe and fine-tuning with $\{4, 8, 16, 32\}$ datapoints from the OOD test set. Our results in Figure 7 show that ensemble reweighting outperforms fine-tuning in the low-data regime (4 and 8) examples, and fine-tuning eventually surpasses the performance of ensemble reweighting.

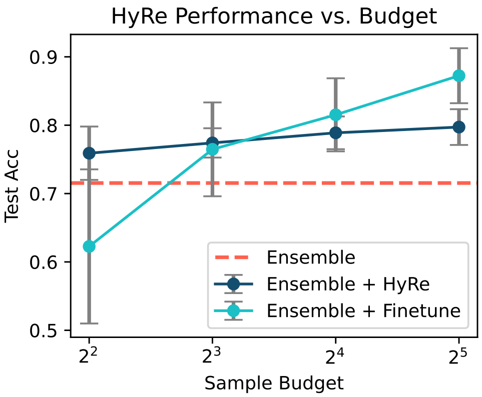


Figure 7: Comparison of HyRe and few-shot fine-tuning on the Camelyon17 OOD test set. HyRe outperforms fine-tuning in the low-data regime despite requiring significantly less computational cost.

B Active Learning Details

We also consider an active learning setup in which the N datapoints to label for HyRe are chosen at test time from a larger unlabeled pool of data. Rather than choosing all datapoints at once, we choose one datapoint at the time based on one of the following three criteria:

- **Entropy** (classification): $H\left(\sum_{h=1}^{H} w_h f_h(x)\right)$. This criterion selects datapoints where the weighted ensemble is most uncertain, promoting the exploration of ambiguous regions.
- **BALD** (classification): $H\left(\sum_{i=1}^{H} w_i f_i(x)\right) - \sum_{i=1}^{H} w_i H(f_i(x))$. BALD considers both ensemble uncertainty and disagreement among members, balancing exploration and exploitation [19, 26].
- **Variance** (regression): $\sum_{i=1}^{H} w_i (f_i(x) - \bar{f}(x))^2$, where $\bar{f}(x) = \sum_{i=1}^{H} w_i f_i(x)$. This criterion focuses on points where ensemble predictions have the highest variance, which is a good indicator of uncertainty in regression tasks.

Each of these criteria can be computed quickly. Because the belief states w has a closed-form update that can be computed very quickly, we can efficiently recompute the next best data point after each active label query.

We note that the first criterion (Entropy) does not distinguish between so-called aleatoric uncertainty and epistemic uncertainty. Therefore, this criterion is susceptible to the "noisy TV problem", where an agent fixates on a source of uncertainty that cannot be resolved [8, 41]. In practice, we find that HyRe is robust to the choice of active learning criterion, and even random selection is effective at adapting to the target distribution.

C Experimental Details

Unless specified otherwise, we use the following configuration for the ensemble networks. We use an ensemble of 100 models. The learnable and prior networks are each a one-hidden-layer MLP with 128 units. For the epinet, the epistemic index is 10-dimensional. For ensemble reweighting via HyRe, we use 32 examples from the target dataset, actively queried based on the BALD (classification) or Variance (regression) criterion. We found that final performance is not very sensitive to the choice of active learning criterion, and even random sampling resulted in consistent benefits.

WILDS. We closely follow the reference WILDS implementation for each dataset [37], including the choice of backbone, learning rate, and weight decay. We briefly describe the baseline methods used in our experiments:

Model	Type	Overall	Chat	Chat Hard	Safety	Reasoning
Mixtral-8x7B-Instruct-v0.1	DPO	77.6	95.0	64.0	72.6	78.7
LLaMA-3-Tulu-2-DPO-70B	DPO	77.2	96.4	57.5	74.9	80.2
Tulu-2-DPO-13B	DPO	76.7	95.8	58.3	79.5	73.2
Tulu-2-DPO-70B	DPO	79.1	97.5	60.5	84.5	74.1
StableLM-2-12B-Chat	DPO	79.9	96.6	55.5	78.1	89.4
Claude-3 Sonnet (June 2024)	Gen	84.2	96.4	74.0	81.6	84.7
GPT-4 (May 2024)	Gen	84.6	96.6	70.4	86.5	84.9
GPT-4 (Aug 2024)	Gen	86.7	96.1	76.1	88.1	86.6
Gemini-1.5-Pro-0924	Gen	86.8	94.1	77.0	85.8	90.2
Skywork-Reward-Gemma-2-27B	Seq	94.3	96.1	89.9	93.0	98.1
INF-ORM-Llama3.1-70B	Seq	95.1	96.6	91.0	93.6	99.1
GRM-Gemma-2B	Seq	84.5	89.4	**75.2**	84.5	88.8
+ Ours (uniform)	Seq	84.5	88.6	72.9	83.7	89.8
+ Ours (N=1)	Seq + HyRe	85.3	88.5	72.7	85.5	91.4
+ Ours (N=5)	Seq + HyRe	86.4	90.3	72.6	89.1	91.4
+ Ours (N=10)	Seq + HyRe	**87.2**	**90.4**	72.5	**90.0**	**92.3**
+ Ours (best head oracle)*	Seq + Oracle	88.6	91.1	78.1	91.9	92.3
+ Ours (best weight oracle)*	Seq + Oracle	90.0	92.3	81.8	92.5	93.1
GRM-Gemma2-2B	Seq	88.4	93.0	**77.2**	92.2	91.2
+ Ours (uniform)	Seq	87.1	96.4	73.1	87.4	89.8
+ Ours (N=1)	Seq + HyRe	86.5	92.4	71.5	85.1	92.5
+ Ours (N=5)	Seq + HyRe	88.5	95.0	72.5	90.3	93.1
+ Ours (N=10)	Seq + HyRe	**89.7**	**96.4**	74.7	**92.4**	**93.5**
+ Ours (best head oracle)*	Seq + Oracle	91.8	97.2	80.0	96.2	94.2
+ Ours (best weight oracle)*	Seq + Oracle	93.1	98.3	83.4	96.7	94.9
Skywork–Llama-3.1-8B	Seq	94.0	94.7	88.6	92.7	96.7
+ Ours (uniform)	Seq	94.0	95.0	87.2	93.0	96.8
+ Ours (N=1)	Seq + HyRe	94.3	95.2	87.8	93.0	97.5
+ Ours (N=5)	Seq + HyRe	94.7	95.5	88.6	93.2	97.8
+ Ours (N=10)	Seq + HyRe	**95.0**	**95.9**	**89.3**	**93.5**	**97.9**
+ Ours (best head oracle)*	Seq + Oracle	96.4	98.3	91.2	95.7	98.4
+ Ours (best weight oracle)*	Seq + Oracle	97.2	99.2	93.0	96.5	98.8

∗ *Oracle* methods show an upper bound on performance, using the test set.

Table 11: Accuracy across tasks in RewardBench. We report overall performance and breakdowns by task category for all models. **HyRe improves upon the state-of-the-art models at the 2B and 8B parameter scales with as few as 1-5 labeled samples per distribution**.

- **CORrelation ALignment** [65, CORAL]: CORAL is an unsupervised domain adaptation method that aligns the second-order statistics (covariances) of source and target feature distributions.
- **Invariant Risk Minimization** [4, IRM]: IRM aims to learn data representations that capture invariant correlations across multiple training distributions.
- **Group Distributionally Robust Optimization** [55, Group DRO]: Group DRO seeks to minimize the worst-case training loss over predefined groups within the data.
- **Fish** [59]: Fish is a domain generalization technique that approximates inter-domain gradient matching by maximizing the inner product between gradients from different domains.
- **LISA** [73]: LISA builds on MixUp and selectively interpolates data samples to achieve domain invariance.

LLM Preference Learning We finetune three reward model checkpoints [72]:

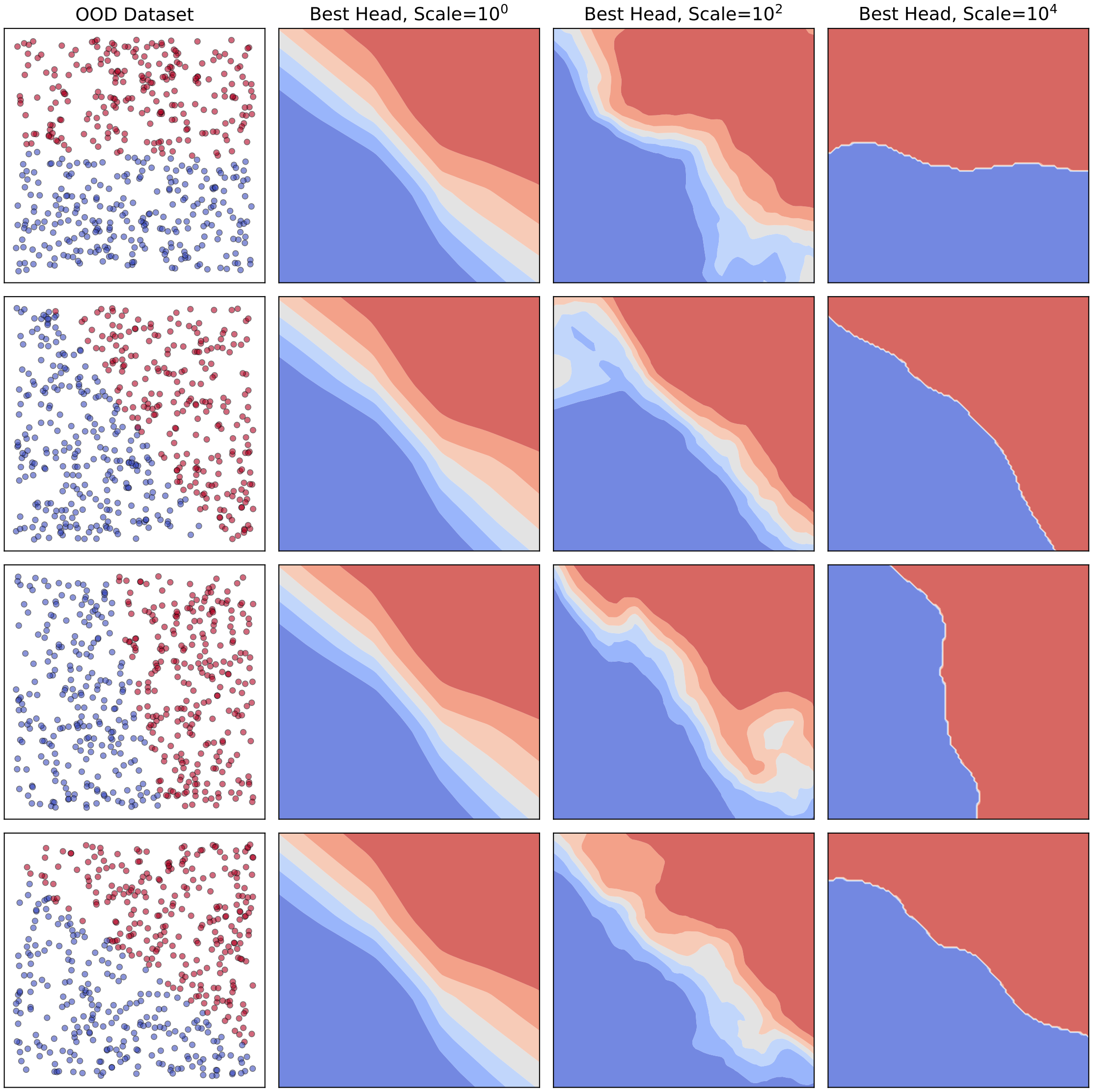


Figure 8: Additional visualizations for the toy conflicting classification example. Increasing the scale hyperparameter results produces heads with sharper decision boundaries.

- https://huggingface.co/Ray2333/GRM-Gemma-2B-rewardmodel-ft
- https://huggingface.co/Ray2333/GRM-Gemma2-2B-rewardmodel-ft
- https://huggingface.co/Skywork/Skywork-Reward-Llama-3.1-8B-v0.2

Our ensemble architecture uses these networks as the backbone, and small MLPs for the learnable and prior networks which take the backbone's final embedding as input. We use the TRL codebase for reward model training [69]. We train with bfloat16 mixed precision. We use a learning rate of 0.0001, no weight decay, a batch size of 16, and train for 5000 steps. We consider four collections of preference datasets:

- **Elix** [60] is inspired by the "Explain like I'm 5" subreddit. It consists of questions answered at five educational levels: elementary, middle, high, college, and expert. Preference pairs are created by scoring how different pairs of GPT-4 generated responses meet the expected comprehension at each level.

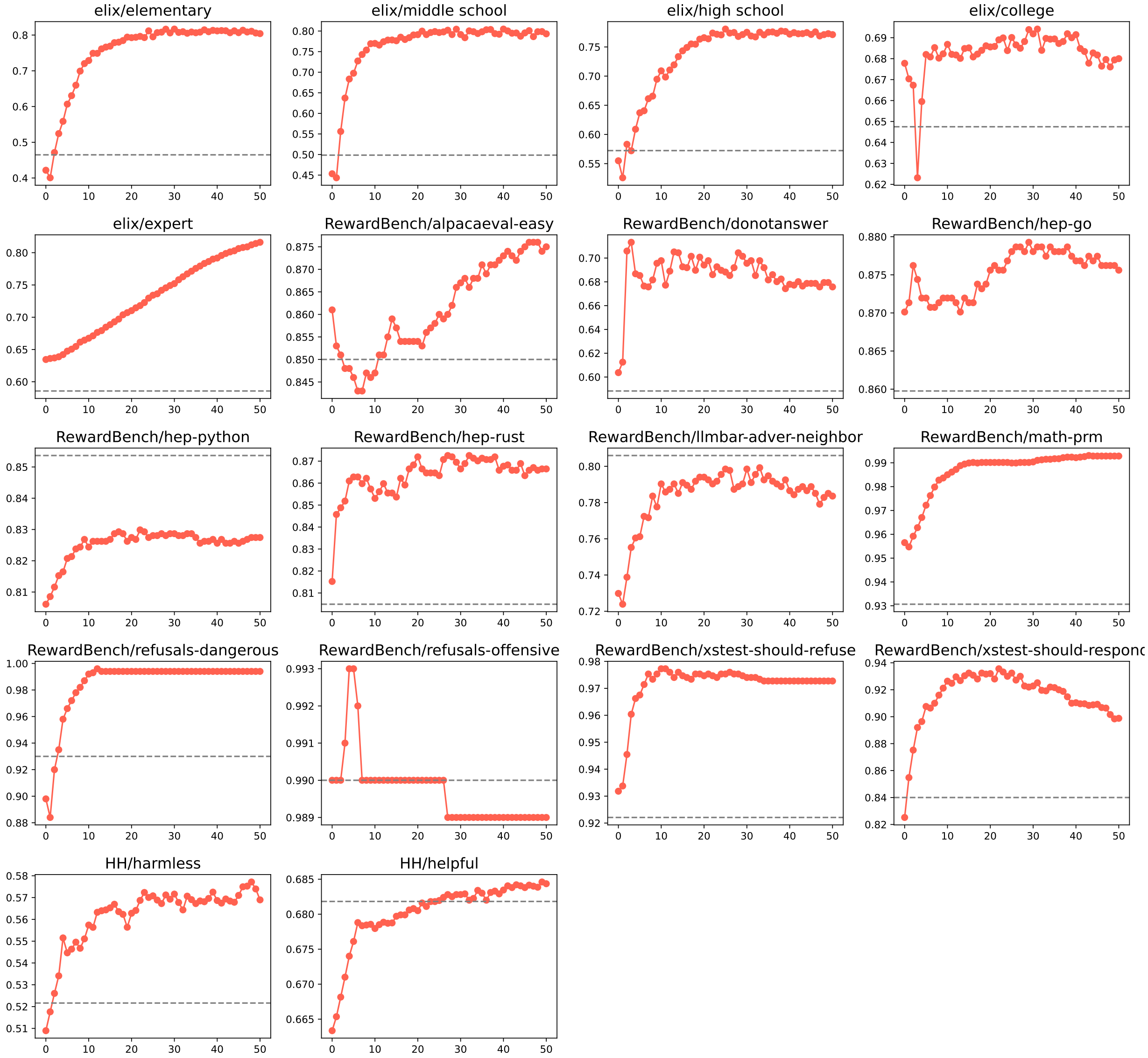


Figure 9: Detailed results for the personalizing preference reward models experiment in Figure 6. Target dataset accuracy (y-axis) after observing different numbers of adaptation samples (x-axis). The dashed line represents the performance of the pretrained reward model.

- **RewardBench** [40] is a suite of 27 preference datasets designed to test reward models on a broad spectrum of tasks, including chat quality, safety, reasoning, coding, and refusal handling. In our aggregate results Figure 6, we drop datasets with less than 100 examples. In our RewardBench experiments Table 11, we use all datasets to ensure a fair comparison with existing methods.

- **PERSONA** [9] contains preference data derived from a collection of synthetic personas with diverse demographic attributes and values. We sample 10 personas and treat each as a target distribution. Further details are in Section G.

- **Anthropic HH** [5] contains human-labeled preferences focused on helpfulness and harmlessness. We use the helpfulness-base and harmlessness-base splits as evaluation distributions to measure the tradeoff between the two objectives.

Algorithm	DL	**CivilComments** Worst-Group Acc	**Amazon** 10% Acc	**FMoW** Worst-Reg Acc	**iWildCam** Macro F1
IRM	O	66.3 (2.1)	52.4 (0.8)	32.8 (2.09)	15.1 (4.9)
IRMX	O	73.4 (1.4)	-	33.7 (0.95)	26.7 (1.1)
IRMX (PAIR)	O	74.2 (1.4)	-	35.4 (1.3)	27.9 (0.9)
CORAL	O	65.6 (1.3)	52.9 (0.8)	32.8 (0.66)	32.7 (0.2)
Group DRO	O	70.0 (2.0)	53.3 (0.0)	31.1 (1.66)	23.8 (2.0)
DFR	O	72.5 (0.9)	-	42.8 (0.42)	-
Fish	O	75.3 (0.6)	53.3 (0.0)	34.6 (0.18)	22.0 (1.8)
LISA	O	72.9 (1.0)	54.7 (0.0)	35.5 (0.81)	-
ERM	X	56.0 (3.6)	53.8 (0.8)	31.3 (0.17)	30.8 (1.3)
Shared-Base	X	58.1 (2.2)	54.2 (0.6)	32.8 (0.4)	30.9 (0.8)
Shared-Base + HyRe	X	58.1 (0.2)	54.2 (0.6)	32.8 (0.4)	31.0 (0.8)

Table 12: Performance on additional WILDS benchmark datasets. The DL column indicates whether the algorithm uses domain labels. Using a Shared-Base ensemble consistently results in gains in OOD generalization metrics over prior methods. However, we observe no further benefits from reweighting the ensemble via HyRe on these datasets.

For the few-shot prompting experiments, we use GPT-4o-mini. For each number of "shots" $N \in \{0, 1, 5, 10, 20, 40, 80\}$, we sample 1000 examples from the target distribution and use them to prompt GPT-4o-mini.

D Diverse Ensemble Architectures

We describe the diverse ensemble architectures used in our experiments. Each architecture is designed to parameterize an ensemble of H models, whose outputs are later combined to form an ensemble prediction. The key goal of these architectures is to produce diverse predictions across the ensemble at a low computational cost.

All architectures are trained end-to-end by minimizing the sum of a standard loss function (cross-entropy for classification, MSE for regression) over all ensemble members:

$$\sum_{h=1}^{H} \mathcal{L}\left(f_h(x), y\right). \tag{4}$$

Here, x is an input example, y is the true label, and f^i is the i-th ensemble member. While each individual model minimizes the training loss, we want the ensemble members to extrapolate to unseen data in diverse ways. The specific ensemble parameterizations, which we describe below, are designed to achieve this goal.

D.1 Vanilla Ensemble

A vanilla ensemble consists of H independently initialized and trained neural networks with identical architectures. Each network f_h takes an input x and produces an output $f_h(x)$. No parameters are shared. While simple to implement, this approach scales poorly as H increases since both memory and computation scale linearly with H.

D.2 Shared-Base Ensemble

We propose a scalable neural network architecture that can represent thousands of diverse ensemble members. The network outputs H real-valued predictions in parallel, with the output space being $\mathbb{R}^H$. The architecture comprises a frozen prior network f_p and a learnable network f_θ, both of which produce outputs of shape $\mathbb{R}^H$. Although the architectures of f_p and f_θ are identical in our experiments, this is not a requirement.

For a given input x, the network output is

$$f^p(z) + f^\theta(z) = \begin{bmatrix} f_1^p(z) + f_1^\theta(z) \\ f_2^p(z) + f_2^\theta(z) \\ \vdots \\ f_H^p(z) + f_H^\theta(z) \end{bmatrix} \in \mathbb{R}^H \tag{5}$$

where each prediction $f_i^p(z) + f_i^\theta(z)$ is compared against the ground-truth label y. The parameters of f^p are fixed at initialization and do not change during training; the parameters of f^θ are learnable.

Using the frozen prior network f^p is crucial to the diversity in this architecture. If we were to only train f^θ, the ensemble of the H predictions would have low diversity due to co-adaptation. To understand why this architecture produces a diverse ensemble, note that each learnable head solves a shifted task determined by the corresponding prior network head. Since we undo this shifting when producing the final prediction, we can view the different learnable heads as solving a different yet equivalent task.

D.3 Epinet

The epinet architecture combines a base model $f^{\text{base}} : \mathcal{X} \to \mathbb{R}^K$ with an epistemic network $f^{\text{epi}} : \mathcal{Z} \times \mathbb{R}^{d_{\text{ftrs}}} \times \mathcal{X} \to \mathbb{R}^K$. The base model can be any regular neural network, including a large pretrained model, and is used to extract features through a feature extractor $\phi : \mathcal{X} \to \mathbb{R}^{d_{\text{ftrs}}}$. Here, d_{ftrs} is the dimension of the extracted intermediate representations.

The epistemic network (epinet) is composed of two parts:

- A frozen prior network $f^{\text{epi-frozen}} : \mathcal{X} \to \mathbb{R}^{1,\dots,d_{\text{index}} \times K}$. The parameters of this network are fixed at initialization and do not change during training.
- A trainable network $f^{\text{epi-trainable}} : \mathcal{Z} \times \mathbb{R}^{d_{\text{ftrs}}} \times \mathcal{X} \to \mathbb{R}^K$.

Given an epistemic index $z \in \mathbb{R}^d$ and input $x \in \mathcal{X}$, we compute the model output as:

$$f(z, x) = f^{\text{base}}(x) + v f^{\text{epi-frozen}}(x) \cdot z + f^{\text{epi-trainable}}(z, \phi(x), x) \cdot z \tag{6}$$

where $\cdot$ is the dot product and $v \in (0, \infty)$ is the so-called prior scale. At each step, we sample multiple epistemic indices z to form an ensemble, i.e., $f_1(x), \dots, f_H(x) = f(z_1, x), \dots, f(z_H, x)$. This architecture efficiently generates diverse predictions by sampling different epistemic indices z while leveraging a potentially large pretrained base model.

E Repulsion vs Random Priors for Diversity

A line of prior work use repulsion for enforcing diversity between ensemble members. The high-level idea is to add a regularization term to the loss function that is minimized when the ensemble members are sufficiently "different" according to some distance metric. For example, Teney et al. [67] uses a repulsion term that maximizes the cosine distance between the gradient of each ensemble member, and Lee et al. [42] maximizes the mutual information of ensemble predictions on OOD inputs. While these techniqueshave seen success in certain settings, our early experiments indicate that such explicit regularization often results in a suboptimal ensemble. The repulsion term can overpower the learning signal in the training data, leading to ensemble members that are diverse but inaccurate.

In contrast, diversification via random priors [49] provides a more balanced approach. The key idea is to initialize each ensemble member with a different random prior function which is fixed throughout training. This introduces diversity from the start without explicitly optimizing for it during training. This approach maintains diversity without sacrificing accuracy on the training data, and the degree of diversification is easily controlled by scaling the prior functions.

F Function-Space Dimensionality Reduction

Here, we expand on the idea of PCA on ensemble predictions. A central challenge with large model ensembles is understanding the commonalities and differences among the individual models. The high-level idea is that PCA applied to ensemble predictions reveals the major direction of variation within an ensemble of models. This dimensionality reduction allows us to clearly interpret model behaviors and identify groups of related datapoints Additionally, PCA enables the generation of new functions with similar statistical properties by parameterizing a low-rank Gaussian distribution in the joint prediction space, which we can sample from.

F.1 Motivating Example

Consider three models $f_1, \ldots, f_3$ and five inputs $z_1, \ldots, z_5$. Denoting each model's predicted probability for an input as $p_{nh} = \sigma(f_h(z_n)) \in [0, 1]$, assume that the matrix of predictions is

$$\begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} \\ p_{21} & p_{22} & p_{23} & p_{24} & p_{25} \\ p_{31} & p_{32} & p_{33} & p_{34} & p_{35} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 & 1/2 \\ 0 & 1 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & 0 & 1 & 1/2 \end{pmatrix}. \tag{7}$$

Each row of this matrix shows one model's prediction on the entire pool of inputs, and each column shows every model's prediction on a single input. We can analyze such a matrix of predictions on three levels, each revealing increasing amounts of structure within the ensemble:

Level 1: Per-sample ensemble uncertainty. We can first compute the average prediction $\bar{p}(x) = \frac{1}{H} \sum_h p_{nh}$ for each datapoint. For the predictions in (7), the average prediction is $\bar{p}(x) = 1/2$ for every input x, and thus the collection of models may be viewed as equally uncertain about each of the 5 inputs. This is the measure of ensemble uncertainty commonly used for ensembles [39].

Level 2: Per-sample disagreement. We can further account for the amount of disagreement among ensemble members for each datapoint. Note that for the four inputs z_1, z_2, z_3, z_4, there is strong disagreement between two functions where one predicts 0 and the other predicts 1. This is not true of z_5, where all functions predict $1/2$. Uncertainty metrics that take disagreement into account, such as the BALD criterion [26], will reveal that the ensemble is more uncertain about z_1, z_2, z_3, z_4 than it is about z_5.

Level 3: Joint predictions. First, note that the two approaches above discard all information about which ensemble member made which individual prediction for a given input, by (1) averaging all predictions or (2) considering only the unordered set of predictions. There is additional structure to the differences among ensemble members that we can extract by considering the joint predictions, i.e., viewing each column of (7) as an object in itself. The pair of inputs (z_1, z_2) are closely related since they deviate from the ensemble prediction in the same "direction" in the joint prediction space ($\mathbb{R}^H$). We can make the same observation about the pair (z_3, z_4). To see this structure more clearly, consider the matrix of deviations from the ensemble prediction $\delta_{nh} = p_{nh} - \frac{1}{H} \sum_h p_{nh}$:

$$\begin{pmatrix} \delta_{11} & \delta_{12} & \delta_{13} & \delta_{14} & \delta_{15} \\ \delta_{21} & \delta_{22} & \delta_{23} & \delta_{24} & \delta_{25} \\ \delta_{31} & \delta_{32} & \delta_{33} & \delta_{34} & \delta_{35} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & -1 & 1 & -1 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \end{pmatrix}. \tag{8}$$

This clearly shows that the vector of joint deviations $(\delta_{11}, \delta_{12}, \delta_{13})$ is the negative of that of $(\delta_{21}, \delta_{22}, \delta_{23})$. More generally, we can view the vector of deviations $(\delta_{1n}, \delta_{2n}, \delta_{3n})$ as a representation of the datapoint z_n in the joint prediction space. In this sense, the matrix of predictions $\{p_{nh}\}$ can be explained by the mean prediction 0.5 for each datapoint, together with two factors of variation $(1, -1, 0)$ and $(1, 0, -1)$ appropriately applied to each input. We next describe how to automatically extract such consistent high-level factors in an ensemble from the matrix of predictions.

F.2 PCA on Ensemble Predictions

We propose to apply PCA to the $H \times N$ matrix of residual predictions to obtain P principal components. Each principle component is a vector of size H that captures the orthogonal factors of variation in how ensemble members extrapolated from the training data. Given a set of weights $w_1, \ldots, w_P$ over principal components, we can "reconstruct" a set of joint predictions as

$$p(x) = \bar{p}(x) + \begin{pmatrix} w_1 & \cdots & w_P \end{pmatrix} \begin{pmatrix} c_{11} & \cdots & c_{1H} \\ c_{21} & \cdots & c_{2H} \\ \vdots & \ddots & \vdots \\ c_{P1} & \cdots & c_{PH} \end{pmatrix} \begin{pmatrix} p_1(x) - \bar{p}(x) \\ p_2(x) - \bar{p}(x) \\ \vdots \\ p_H(x) - \bar{p}(x) \end{pmatrix}, \tag{9}$$

where we denote the mean prediction as $\bar{p}(x) = \frac{1}{H} \sum_h p_{nh}$ and the P principal components as $C \in \mathbb{R}^{P \times H}$.

We highlight two known interpretations of PCA that have interesting implications for our goal of summarizing ensemble predictions:

Maximum mutual information / variance after projection. PCA finds the linear projection $y = w^\top x$ with unit vector w that achieves maximum mutual information $I(x; y)$, or equivalently, maximum variance $\mathrm{Var}(y)$. Each principal component finds the linear combination of ensemble members that preserves the most information about the set of joint ensemble predictions. This is closely related to the disagreement term in Bayesian active learning [26].

Factor model. The principal components are maximum likelihood parameters under a linear Gaussian factor model of the data [68]. Indeed, we can view our principal components as orthogonal modifications to the mean prediction $\bar{p}(x)$. The distribution of ensemble members is closely approximated by "reconstructed predictions" (9), where $z_{1:P} \sim \mathcal{N}(0, \mathrm{I}^P)$. We can view each principal component as a consistent high-level direction of functional variation in which the training data provided insufficient information.

G PERSONA Dataset Details

Below, we list the personas used in our PERSONA [9] experiments. The dataset includes 1000 personas in total, each with 200 preference pairs. We subsampled 10 personas from the original dataset of 1000, ensuring a diverse set of backgrounds, ages, and lifestyles.

Persona 1. Age: 1. Sex: Male. Race: White alone. Ancestry: Irish. Household language: English only. Education: Not applicable. Employment status: Not applicable. Class of worker: Not applicable. Industry category: Not applicable. Occupation category: Not applicable. Detailed job description: Not applicable. Income: Not applicable. Marital status: Too young to be married. Household type: Cohabiting couple household with children of the householder less than 18. Family presence and age: With related children under 5 years only. Place of birth: Missouri/MO. Citizenship: Born in the United States. Veteran status: Not applicable. Disability: None. Health insurance: With health insurance coverage. Fertility: Not applicable. Hearing difficulty: None. Vision difficulty: None. Cognitive difficulty: None. Ability to speak english: Not applicable. Big five scores: Openness: High, Conscientiousness: High, Extraversion: Low, Agreeableness: Extremely High, Neuroticism: Extremely Low. Defining quirks: Loves to play with his food. Mannerisms: Waves hands when excited. Personal time: Spends most of his time playing, sleeping, and learning to walk. Lifestyle: Lives a carefree and playful lifestyle. Ideology: Not applicable. Political views: Not applicable. Religion: Other Christian.

Persona 2. Age: 11. Sex: Male. Race: White alone. Ancestry: Irish. Household language: English only. Education: Grade 4. Employment status: Unemployed. Class of worker: Not applicable. Industry category: Not applicable. occupation category: Not applicable Detailed job description: Student. Income: 0. Marital status: Never married or under 15 years old. Household type: Cohabiting couple household with children of the householder less than 18. Family presence and age: With related children 5 to 17 years only. Place of birth: Louisiana/LA. Citizenship: Born in the United States. Veteran status: Not applicable. Disability: None. Health insurance: With health insurance coverage. Big five scores: Openness: Low, Conscientiousness: Low, Extraversion: High, Agreeableness: High, Neuroticism: Average. defining quirks: Loves to draw and create stories Mannerisms: Often seen doodling or daydreaming. Personal time: Spends free time drawing or playing video games. Lifestyle: Active and playful, enjoys school and spending time with friends. Ideology: Undeveloped. Political views: Undeveloped. Religion: Religiously Unaffiliated.

Persona 3. Age: 19. Sex: Male. Race: Asian Indian alone. Ancestry: Indian. Household language: Hindi. Education: 1 or more years of college credit, no degree. Employment status: Not in labor force. Class of worker: Not Applicable. Industry category: Not Applicable. Occupation category: Not Applicable. Detailed job description: Not Applicable. Income: -60000.0. Marital status: Never married or under 15 years old. Household type: Living with parents. Family presence and age: Living with two parents. Place of birth: India. Citizenship: Not a U.S. citizen. Veteran status: Non-Veteran. Disability: None. Health insurance: With health insurance coverage. Big five scores: Openness: Average, Conscientiousness: High, Extraversion: Extremely Low, Agreeableness: Extremely High, Neuroticism: Extremely Low. defining quirks: Passionate about music Mannerisms: Expressive hand gestures when speaking. Personal time: Practicing music or studying. Lifestyle: Student and Music Enthusiast. Ideology: Liberal. Political views: Liberal. Religion: Other Christian.

Persona 4. Age: 29. Sex: Female. Race: Laotian alone. Ancestry: Laotian. Household language: Asian and Pacific Island languages. Education: Some college, but less than 1 year. Employment status: Armed forces, at work. Class of worker: Federal government employee. Industry category: MIL-U.S. Navy. Occupation category: MIL-Military Enlisted Tactical Operations And Air/Weapons Specialists And Crew Members. Detailed job description: Maintains and operates tactical weapons systems. Income: 81000.0. Marital status: Married. Household type: Married couple household with children of the householder less than 18. Family presence and age: With related children 5 to 17 years only. Place of birth: California/CA. Citizenship: Born in the United States. Veteran status: Now on active duty. Disability: None. Health insurance: With health insurance coverage. Big five scores: Openness: Average, Conscientiousness: High, Extraversion: Average, Agreeableness: High, Neuroticism: Average. Defining quirks: Collects military memorabilia. Mannerisms: Frequently uses military jargon. Personal time: Spends time with family and collecting military memorabilia. Lifestyle: Disciplined and active. Ideology: Conservative. Political views: Republican. Religion: Protestant.

Persona 5. Age: 36. Sex: Female. Race: Some Other Race alone. Ancestry: Hispanic. Household language: English. Education: Regular high school diploma. Employment status: Civilian employed, at work. Class of worker: Employee of a private for-profit company or business, or of an individual, for wages, salary, or commissions. Industry category: FIN-Insurance Carriers. Occupation category: OFF-Insurance Claims And Policy Processing Clerks. Detailed job description: Processes insurance claims and policies. Income: 182000.0. Marital status: Married. Household type: Married couple household with children of the householder less than 18. Family presence and age: With related children under 5 years only. Place of birth: New Mexico/NM. Citizenship: Born in the United States. veteran status: Non-Veteran Disability: None. Health insurance: With health insurance coverage. Big five scores: Openness: Extremely Low, Conscientiousness: Extremely High, Extraversion: Extremely High, Agreeableness: High, Neuroticism: Average. Defining quirks: Enjoys bird-watching. Mannerisms: Often taps foot when thinking. Personal time: Spends free time with family or in nature. Lifestyle: Active and family-oriented. Ideology: Conservative. Political views: Republican. Religion: Other Christian.

Persona 6. Age: 44. Sex: Female. Race: Black or African American alone. Ancestry: Haitian. household language: Other Indo-European languages education: Associate's degree Employment status: Civilian employed, at work. Class of worker: Employee of a private not-for-profit, tax-exempt, or charitable organization. Industry category: FIN-Banking And Related Activities. Occupation category: OFF-Tellers. Detailed job description: Handles customer transactions at the bank, including deposits, withdrawals, and loan payments. Income: 40000.0. Marital status: Separated. Household type: Female householder, no spouse/partner present, with children of the householder less than 18. Family presence and age: With related children 5 to 17 years only. Place of birth: Haiti. Citizenship: Not a U.S. citizen. Veteran status: Non-Veteran. Disability: None. Health insurance: With health insurance coverage. Big five scores: Openness: High, Conscientiousness: Extremely Low, Extraversion: Average, Agreeableness: Average, Neuroticism: Extremely Low. Defining quirks: Loves to cook Haitian cuisine. Mannerisms: Often taps her foot when stressed. Personal time: Taking care of her children, Pursuing further education. Lifestyle: Busy, Family-oriented. Ideology: Egalitarian. Political views: Democrat. Religion: Protestant.

Persona 7. Age: 52. Sex: Female. Race: Korean alone. Ancestry: Korean. Household language: Asian and Pacific Island languages. Education: Regular high school diploma. Employment status: Civilian employed, at work. Class of worker: State government employee. Industry category: ENT-Restaurants And

Other Food Services. Occupation category: EAT-First-Line Supervisors Of Food Preparation And Serving Workers. Detailed job description: Supervises food preparation and serving workers in a state government facility. Income: 133900.0. Marital status: Married. Household type: Married couple household, no children of the householder less than 18. Family presence and age: No related children. Place of birth: Korea. Citizenship: U.S. citizen by naturalization. Veteran status: Non-Veteran. Disability: None. Health insurance: With health insurance coverage. big five scores: Openness: Average, Conscientiousness: Extremely High, Extraversion: Extremely Low, Agreeableness: Extremely Low, Neuroticism: Average defining quirks: Deep love for literature and reading Mannerisms: Constantly adjusts her glasses. Personal time: Spends free time reading or engaging in community activism. Lifestyle: Quiet and community-oriented. Ideology: Liberal. Political views: Democratic. Religion: Protestant.

Persona 8. Age: 58. Sex: Male. Race: White. Ancestry: Scottish. Household language: English. Education: Bachelor's Degree. Employment status: Employed. Class of worker: Private. industry category: Investigation And Security Services Occupation category: Sales Manager. Detailed job description: Oversees sales teams, sets sales goals, and develops strategies to achieve these goals. Income: 198200. Marital status: Married. Household type: Married couple household, no children under 18. Family presence and age: No related children. Place of birth: Florida. Citizenship: US Citizen. veteran status: Non-Veteran Disability: With a disability. Health insurance: With health insurance coverage. Big five scores: Openness: High, Conscientiousness: Extremely High, Extraversion: Average, Agreeableness: Average, Neuroticism: Average. Defining quirks: Keen interest in security technology and crime novels. mannerisms: Constantly checks his surroundings Personal time: Researching the latest security technologies or enjoying a round of golf. Lifestyle: Active and health-conscious. Ideology: Conservative. Political views: Republican. Religion: Catholic.

Persona 9. Age: 65. Sex: Female. Race: White alone. Ancestry: Italian. Household language: Other Indo-European languages. Education: Master's degree. Employment status: Civilian employed, at work. Class of worker: Self-employed in own incorporated business, professional practice or farm. Industry category: ENT-Traveler Accommodation. Occupation category: FIN-Accountants And Auditors. Detailed job description: Manages financial records and tax data for her own travel accommodation business. Income: 188600.0. Marital status: Married. Household type: Married couple household, no children of the householder less than 18. Family presence and age: No related children. Place of birth: Delaware/DE. Citizenship: Born in the United States. Veteran status: Non-veteran. Disability: None. Health insurance: With health insurance coverage. ability to speak english: Well. Big five scores: Openness: Average, Conscientiousness: Low, Extraversion: Low, Agreeableness: Average, Neuroticism: Extremely High. Defining quirks: Has an extensive collection of vintage travel posters. Mannerisms: Tends to use Italian phrases in conversation. Personal time: Spends her free time exploring new places, trying new cuisines, and learning about different cultures. Lifestyle: Leads a busy lifestyle managing her business, but always finds time for her passion for travel and culture. Ideology: Believes in the importance of understanding and appreciating different cultures. Political views: Liberal. Religion: Protestant.

Persona 10. Age: 75. Sex: Female. Race: White alone. ancestry: Scottish Household language: English only. Education: Professional degree beyond a bachelor's degree. Employment status: Not in labor force. Class of worker: Retired. Industry category: Healthcare. Occupation category: Doctor. Detailed job description: Retired pediatrician. Income: 98000.0. Marital status: Never married. Household type: Female householder, no spouse/partner present, living alone. Family presence and age: No family. Place of birth: Massachusetts/MA. citizenship: Born in the United States veteran status: Non-Veteran Disability: None. Health insurance: With health insurance coverage. Big five scores: Openness: Average, Conscientiousness: Average, Extraversion: High, Agreeableness: Extremely High, Neuroticism: Average. Defining quirks: Enjoys cooking traditional Scottish meals. Mannerisms: Often hums traditional Scottish tunes. Personal time: Spends free time volunteering at the local church and community center. Lifestyle: Active but relaxed, with a focus on maintaining health and staying involved in the community. Ideology: Conservative. Political views: Republican. Religion: Catholic.