TEST-TIME ALIGNMENT VIA HYPOTHESIS REWEIGHTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large pretrained models often struggle with *underspecified* tasks—situations where the training data does not fully define the desired behavior. For example, chatbots must handle diverse and often conflicting user preferences, requiring adaptability to various user needs. We propose a novel framework to address the general challenge of aligning models to test-time user intent, which is rarely fully specified during training. Our approach involves training an efficient ensemble, i.e., a single neural network with multiple prediction heads, each representing a different function consistent with the training data. Our main contribution is HYRE, a simple adaptation technique that dynamically reweights ensemble members at test time using a small set of labeled examples from the target distribution, which can be labeled in advance or actively queried from a larger unlabeled pool. By leveraging recent advances in scalable ensemble training, our method scales to large pretrained models, with computational costs comparable to fine-tuning a single model. We empirically validate HYRE in several underspecified scenarios, including personalization tasks and settings with distribution shifts. Additionally, with just five preference pairs from each target distribution, the same ensemble adapted via HYRE outperforms the prior state-of-the-art 2B-parameter reward model accuracy across 18 evaluation distributions.

1 INTRODUCTION

Task specification—the process of communicating the desired behavior to a machine learning model—is inherently iterative and rarely complete after a finite set of instructions or training examples. Addressing task underspecification is a fundamental challenge in machine learning, especially as models are employed for increasingly complex and nuanced tasks. For example, personalizing a chatbot assistant is difficult because chatbots are typically trained to optimize an aggregate preference metric through Reinforcement Learning from Human Feedback (RLHF) (Siththaranjan et al., 2023), using preference labels collected from a diverse set of users. This often leads to responses that do not align with individual user needs since different users have conflicting preferences shaped by individual backgrounds and experiences. The main challenge lies in adapting the model's behavior to suit each user based on minimal additional input. Underspecification can also arise due to other factors, such as spurious correlations in the training data, insufficient training samples, label noise, and limitations in supervisor ability. Our broader goal is to leverage the diverse latent capabilities inside large pretrained models to facilitate adaptation with minimal additional supervision.

Existing methods for adapting models to previously underspecified tasks generally fall into two categories: (1) optimizing zero-shot inputs, such as natural language prompts (Gao et al., 2020; Khattab et al., 2023; Yuksekgonul et al., 2024), or (2) fine-tuning the model's parameters (Houlsby et al., 2019; Hu et al., 2021; Liu et al., 2024; Wu et al., 2024). While recent works have made progress on both fronts, these approaches remain insufficient for real-time adaptation. Specifically, these prior methods require substantial computational resources, involving at the very least multiple passes through the model. Moreover, they are "passive" in nature, not allowing models to actively request additional information to resolve ambiguities. As a result, these approaches are impractical for on-the-fly adaptation at test time, where quick responses are critical.

To address these limitations, we build on recent advances in efficient ensemble architectures, which enable a single neural network to represent a collection of diverse models with minimal computational overhead (Osband et al., 2023). These architectures naturally model task ambiguity as a set

of possible functions, enabling one network to capture several plausible interpretations of the training data. By dynamically switching between these functions based on how well each performs on target inputs, the model can better align with the user's intent. While previous work has used such architectures to quantify uncertainty, we propose to use them to resolve ambiguity at test time. To achieve this, we develop a method that efficiently reweights a given ensemble of models based on a few labeled target examples.

We propose Hypothesis Reweighting (HyRe), a simple and computationally efficient method for test-time task disambiguation. Our approach consists of two steps. First, we train a diverse ensemble of models on training data, with each model initialized from the same pretrained backbone. Then, at test time, we evaluate the performance of each ensemble member on a small set of examples from the target distribution, which can be labeled in advance or actively queried from a larger unlabeled pool. Based on their performance, we dynamically reweight the ensemble, assigning higher weights to models that are more aligned with the target distribution. This reweighted ensemble is our final model, which we use for making predictions on new, unseen data. HyRe is an instance of generalized Bayesian inference (Bissiri et al., 2016): starting from a maximum entropy prior (i.e., a uniformly weighted ensemble), the procedure converges to the optimal weighting over ensemble members given sufficient i.i.d. examples from the target distribution. To our best knowledge, HyRe is the first to apply this framework to adapting deep network ensembles using non-differentiable performance metrics such as 0-1 error.

We evaluate HyRe using two ensemble architectures across over 20 target distributions, spanning WILDS distribution shifts, preference personalization tasks, and benchmarks for safety and usefulness in responses. Our findings show that HyRe enables rapid test-time adaptation of large models with minimal targeted feedback, requiring as few as five labeled examples. Notably, in a preference personalization setting with 5 distinct evaluation personas, HyRe achieves an average 20% accuracy gain over the previous state-of-the-art model at 2B parameter scale. These results demonstrate that HyRe can effectively resolve task underspecification with minimal labeled data at test time.

2 Preliminaries

2.1 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. The training distribution is denoted by $\mathrm{P}_{\text{train}}$, and the evaluation distribution by P_{eval}, both 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, where each pair (x_i, y_i) is drawn from $\mathrm{P}_{\text{train}}$. We explore several underspecified settings—scenarios where data drawn from $\mathrm{P}_{\text{train}}$ only partially informs the model on performing under P_{eval}. For instance, in a chatbot personalization task, the training data from $\mathrm{P}_{\text{train}}$ may include a variety of user preferences regarding response styles, while the test distribution P_{eval} might involve a specific new user with unique preferences.

To enable the model to quickly improve its performance under P_{eval} at test time, we give it access to a small adaptation dataset $\mathcal{D}_{\text{adapt}} \sim \mathrm{P}_{\text{eval}}$. The adaptation data can be labeled in advance (few-shot learning) or actively queried from a pool of unlabeled data (active learning). This dataset is significantly smaller than the training dataset ($|\mathcal{D}_{\text{adapt}}| \ll |\mathcal{D}_{\text{train}}|$) and is intended for on-the-fly adaptation without further model training. As a point of reference, in our main experiment, we have $|\mathcal{D}_{\text{adapt}}| = 16$ and $|\mathcal{D}_{\text{train}}| > 300,000$, with the adaptation step occurring near-instantly after passing $\mathcal{D}_{\text{adapt}}$ through the network once.

2.2 Efficient Ensembles

We train an ensemble of K models $f_1, \ldots, f_K$ on the training data $\mathcal{D}_{\text{train}}$. We consider paramtererizations 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* (Osband et al., 2023), which are 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. Specifically, prior networks prevent *ensemble collapse* when using a shared backbone—a scenario where all ensemble members converge to similar functions even outside the training distribution, offering no advantage over a single model. While we expect ensemble members to produce nearly identical

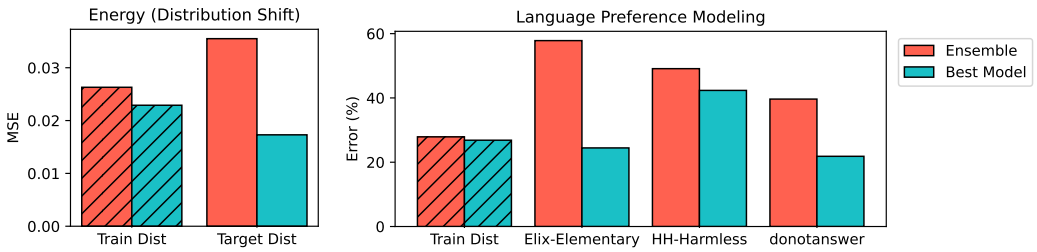

Figure 1: **The ensemble average is suboptimal in underspecified tasks.** Performance of the uniform ensemble vs. the best individual model across four underspecified tasks (lower is better). In all cases, the best single head outperforms the uniform ensemble on the target distribution, highlighting the need for approaches that utilize additional information about the target distribution to optimize ensemble weighting.

predictions on the training data (i.e. achieving low training loss), we aim for their predictions on unseen target data to remain diverse, reflecting the range of functions consistent with the training data. In our experiments, we consider two ensemble architectures which are designed to scalably represent an ensemble of models:

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 the task loss of each model $\sum_{k=1}^{K} \mathcal{L}(f_k, \mathcal{D}_{\text{train}})$ using stochastic gradient descent. We emphasize that these architectures are computationally efficient, and have negligible overhead compared to a single model. In our reward model experiments, for instance, 100 ensemble heads add only 550 thousand parameters (0.03%) to the 2 billion parameter Gemma backbone. Please refer to Appendix C for a more detailed description of these ensemble architectures. In the next section, we propose a simple method for efficiently leveraging ensemble diversity to quickly adapt to new data distributions.

3 Test-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 reweights an ensemble of models, leveraging their diversity to prioritize functions most consistent with new test-time data. This allows us to efficiently improve performance in scenarios where the task is underspecified, and additional data can help resolve ambiguity. The adaptation happens entirely during test time and makes no change to model parameters (as done for fine-tuning).

3.1 The Ensemble Average is Suboptimal in Underspecified Tasks

Conventional wisdom suggests that an ensemble of independently trained models outperforms each individual model by averaging out errors, a principle widely supported for improving performance and uncertainty estimation (Krogh & Vedelsby, 1994; Hansen & Salamon, 1990; Dietterich, 2000; Lakshminarayanan et al., 2017; Ovadia et al., 2019). The (uniform) ensemble average leverages the diversity among models to reduce variance and mitigate individual model biases.

However, recent works have shown that in highly underspecified conditions—where the training data does not sufficiently define the desired behavior on target inputs—the ensemble average can be suboptimal (Teney et al., 2022; Lee et al., 2023). In such settings, ensemble members may adopt different implicit assumptions to address gaps in the training data, resulting in a diverse set of internally consistent functions. Uniformly averaging these functions effectively "blends" their underlying assumptions, leading to outputs that may not align well with specific target behavior. Instead, dynamically reweighting the ensemble based on alignment with the target task can better capture the desired behavior and improve performance.

These earlier observations focused on synthetic tasks designed to highlight the shortcomings of point estimates. We seek to evaluate whether these findings extend to large-scale real-world scenarios where the underspecification is more subtle, such as in distribution shifts and personalization tasks. Our results in Figure 1 confirm that the ensemble average is indeed suboptimal in these settings, achieving lower performance than a single model in the ensemble.

Algorithm 1 HYRE: Test-Time Hypothesis Reweighting

Require: Ensemble members $f_{1..H}$, unlabeled dataset $x_{1..N}$, query budget B, prior weight p
1: Initialize weights $w \leftarrow [\frac{1}{H}, \ldots, \frac{1}{H}]$, query set $Q \leftarrow \emptyset$
2: **for** $i \leftarrow 1$ to B **do**
3: (Optional) Query label y_n for $\arg\max_n c(x_n)$ and add (x_n, y_n) to Q (Appendix A)
4: Compute accuracy $\text{acc}_h = \sum_{n \in Q} \text{acc}(f_h, x_n, y_n)$ for each h
5: Update ensemble weight $w_h \propto \exp(\text{acc}_h + p)$ (Section 3.2)
6: **end for**
7: **Return** final weighted ensemble function $f_w : x \mapsto \sum_{h=1}^{H} w_h f_h(x)$

Building on this empirical observation, we propose to adjust the weights assigned to each ensemble member based on their alignment with the target task. By leveraging a small amount of labeled target data, we can reweight the ensemble to emphasize models whose implicit assumptions best match the given target task. In Section 3.2, we describe our method for dynamically reweighting ensemble members, which finds an ensemble weighting that better aligns with a given target distribution. We validate the effectiveness and sample efficiency of this approach through comprehensive experiments in Section 6.

3.2 FAST ENSEMBLE REWEIGHTING

Given an ensemble of K models $f_1, \ldots, f_K$, we aim to dynamically update the weights assigned to each model based on additional data. As a practical test-time assumption in settings where we cannot further train neural networks, we can think of the "true" or "best" model as being one of the H ensemble particles that performs best on the evaluation distribution. Initially, we assign equal weights $w_h = \frac{1}{H}$ to all ensemble members to reflect a uniform prior belief over the ensemble members. As new labeled data becomes available, we update w according to which model is most appropriate for P_{eval}.

Formally, the weighted ensemble prediction is computed as $f_w(x) = \sum_{i=1}^{K} w_i f_i(x)$, where w_i are nonnegative weights satisfying $\sum_{i=1}^{K} w_i = 1$. To update the weights in light of new adaptation data, we use an objective function $l(f_k, x, y)$ that is used to measure the performance of each ensemble member f_k on the datapoint (x, y). We define the cumulative loss of model f_k on the adaptation data as

$$\mathcal{L}(f_k, \mathcal{D}_{\text{adapt}}) = \sum_{(x,y) \in \mathcal{D}_{\text{adapt}}} l(f_k, x, y). \tag{1}$$

We then compute updated weights using a softmax on the negative cumulative losses:

$$w_h = \frac{\exp(-\mathcal{L}(f_h, \mathcal{D}_{\text{adapt}}))}{\sum_{k=1}^{H} \exp(-\mathcal{L}(f_k, \mathcal{D}_{\text{adapt}}))}. \tag{2}$$

Here, the weights w_h sum to 1, assigning greater weight to models that perform well on adaptation data. In our experiments, we use the 0-1 error for classification and the mean squared error for regression as the objective function $l(f_k, x, y)$.

As an optional assumption for further performance gains, we also consider an active learning setup in which the N datapoints to label are chosen at test time from a larger unlabeled pool of data. We summarize the overall fast adaptation procedure in Algorithm 1, and describe the active learning setup in Appendix A. HYRE is particularly well-suited for real-world deployment settings, where fast adaptation or personalization is required within seconds. Compared to fine-tuning the model with SGD, this approach is significantly faster and even outperforms it in the low-data regime, as we will see in Section 4. Our experiments will explore these adaptation capabilities across a variety of tasks and datasets. The choice of active learning criterion further allows optimization for specific task requirements, making this a versatile framework for addressing a wide range of real-world challenges.

3.3 INTERPRETATION AS GENERALIZED BAYESIAN INFERENCE.

The weight update in (2) can be interpreted as a form of generalized Bayesian inference (Bissiri et al., 2016). Given an initial belief state $\pi(w)$, the updated belief after observing $\mathcal{D}_{\text{adapt}}$ is:

$$\pi(w|\mathcal{D}_{\text{adapt}}) \propto \exp\left(-\mathcal{L}(w, \mathcal{D}_{\text{adapt}})\right) \pi(w), \tag{3}$$

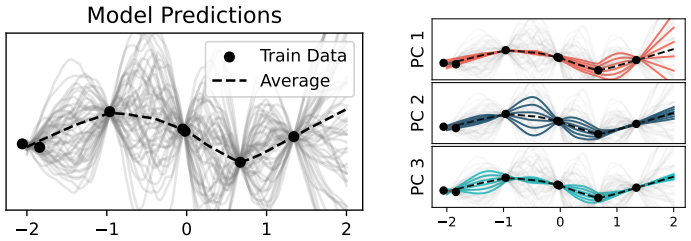

Figure 2: Principal component analysis of an ensemble of regression models. (Left) The ensemble of functions, with each gray line representing one function. The dashed line shows the (average) ensemble prediction. (Right) The first three principal components of the ensemble's predictions. Each principal component reflects a distinct functional variation.

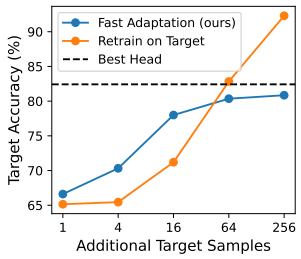

Figure 3: Performance of HyRe vs fine-tuning at different amounts of adaptation data. Ensemble reweighting outperforms fine-tuning in the low-data regime.

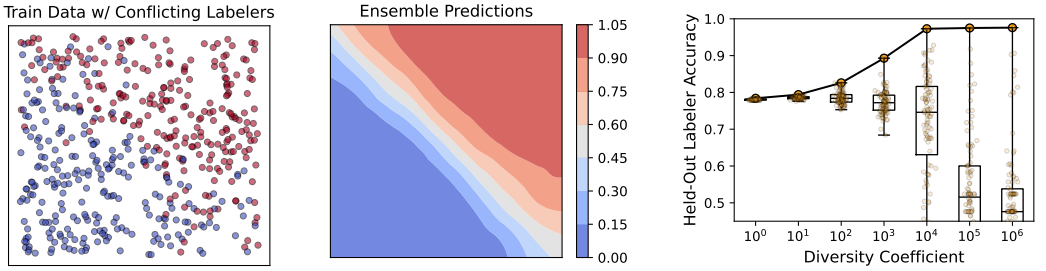

Figure 4: Visualization of an ensemble model trained on data with conflicting labels. (Left) The training dataset is labeled by multiple labelers with conflicting preferences, introducing ambiguity. (Center) The average predictions of an ensemble capture the "average labeler", resulting in smooth decision boundaries that blend the conflicting input. (Right) Increasing diversity leads to a population with higher maximum agreement with a held-out labeler.

This framework generalizes classical Bayesian inference by allowing arbitrary loss functions $l(w, x)$, providing more flexibility than the traditional likelihood approach. Note that we recover standard Bayesian inference when $l(w, x) = -\log p(x|w)$ (i.e., the log likelihood). As shown in Bissiri et al. (2016), this flexibility is particularly useful when the underlying model is misspecified, as is the case for underspecified tasks. Importantly, updates of this form are both consistent and coherent, ensuring that this belief update is the only rational one based on the initial belief state and observed data. Specifically, such updates converge to the true parameters as the amount of data increases (consistency), and successive updates, whether from a single observation or accumulated over multiple observations, lead to the same posterior (coherence).

In classification tasks, using log-likelihood as $l(w, x)$ often leads to belief updates being dominated by outliers or extreme datapoints. By contrast, employing alternative loss functions such as the 0-1 error ensures consistent scaling of $l(w, x)$ values and results in more stable updates to w. This is in line with observations in Izmailov et al. (2021), which found that relying on log-likelihood can cause issues on inputs outside the training distribution. Thus, generalized Bayesian inference allows for more robust belief updates in real-world settings.

4 When is Ensemble Reweighting Effective, and Why?

In Section 3, we introduced fast adaptation and argued that it is particularly effective when the desired test-time behavior is underspecified. This section further explores this hypothesis through illustrative examples. First, we demonstrate a simple method using PCA to analyze differences between ensemble members, revealing that the ensemble serves as a nonparametric representation of task ambiguity. We then show that diverse ensembles can uncover distinct, sharp decision boundaries, which together explain aggregate behavior while capturing conflicting labeler preferences. Finally, we explore the tradeoffs between fast adaptation and fine-tuning, demonstrating that fast adaptation is more advantageous when target data is limited.

The differences between ensemble members reflect task ambiguity. We explore the extent to which a diverse ensemble can serve as a nonparametric representation of task ambiguity. We consider a synthetic regression task where the training data is sampled from a Gaussian Process (GP) prior, and the goal is to adapt to one of several possible functions sampled from the GP poste-

rior conditioned on the training data. Each ensemble member f_k produces a vector of predictions $v_k \in \mathbb{R}^M$ at a set of target inputs $x_1, \ldots, x_M$. By performing Principal Component Analysis (PCA) on the matrix of predictions $V = (v_1, \ldots, v_K) \in \mathbb{R}^{M \times K}$, we extract principal components $u_1, \ldots, u_m \in \mathbb{R}^M$ that summarize the primary modes of variation variation between ensemble members. Each prediction vector v_k can be approximated as a weighted sum of the principal components, i.e., $v_k \approx \sum_{i=1}^{m} \alpha_{k,i} u_i$ for some coefficients $\alpha_{k,i}$.

We train an ensemble of 100 models on a dataset of 7 inputs, and evaluate on a held-out set of 1000 test inputs. We visualize the first three principal components extracted from the ensemble in Figure 2. Each principal component reflects a distinct mode of variation, reflecting different local function variations while maintaining smoothness and fit to the training data. These components may be seen as similar to wavelets, in that the most of the variation from one principle component is "local" in input space, and these components form a basis that can approximate the ensemble. We refer the interested reader to Appendix E for further motivation and intuition for PCA applied to the ensemble predictions.

Diverse ensembles uncover many sharp decision boundaries. The Bradley-Terry model is commonly used to describe pairwise comparisons between items, using latent parameters to represent each item's quality. For two items i and j with latent parameters $\theta_i, \theta_j \in \mathbb{R}$, the probability of i being preferred over j is given by $P(i \succ j) = \frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j}}$. While this model is often interpreted as describing one stochastic decision maker, it can equivalently be interpreted as describing a pool of deterministic decision makers. To see this, it is helpful to consider the following equivalent form of Bradley-Terry:

$$P(i \succ j) = \frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j}} = P\left(\theta_i + \epsilon_i > \theta_j + \epsilon_j\right) \quad \text{where} \quad \epsilon_i, \epsilon_j \sim \text{Gumbel}(0, 1). \tag{4}$$

To generate a label $i \succ j$, we first sample a decision maker represented by the pair (ϵ_i, ϵ_j). Given the pair (ϵ_i, ϵ_j) the choice between i and j is deterministic. By averaging over many such deterministic decisions, each influenced by different realizations of ϵ_i, ϵ_j, we recover the probabilistic preference described by the original Bradley-Terry model. In this sense, the model can be reinterpreted as an ensemble of deterministic decision-makers with different biases. Each decision-maker has fixed preferences for any given pair, but across the population, these preferences reflect the overall probabilistic distribution.

This equivalance allows us to see the Bradley-Terry model as an aggregate of deterministic decision-makers rather than a purely stochastic process, with randomness introduced through individual biases (represented by ϵ_i, ϵ_j). This perspective suggests that if the preference labels are generated by a pool of underlying functions or annoators then it may be helpful to identify the decision boundaries associated with each annotator; if we do so, we can "personalize" the model at test time by quickly figuring out which ensemble member best describe the person at hand. We hypothesize that ensemble architectures with priors towards diversity naturally encourage the learning of such diverse decision makers.

To test this hypothesis, we consider an illustrative contextual preference learning task with conflicting labelers. We randomly sample inputs (x_1, x_2) from $[0, 1]^2$, and simulate a diverse set of deterministic decision-makers, each corresponding to a different linear decision boundary in 2D input space. Specifically, we sample $w_1, w_2 \sim N(0, 1)$ and set the decision boundary to be $w_1 x_1 + w_2 x_2 > 0$. After training a diverse ensemble model on this dataset, we evaluate its performance on different decision boundaries. Results in Figure 4 show that the ensemble can quickly adapt to a new decision boundary, outperforming a single model on the decision boundary it was trained on. We find that the average ensemble prediction matches the "average decision maker", while each individual ensemble member corresponds to a sharp decision boundary. Using a lower temperature during training results in sharper decision boundaries for each ensemble member. In Section 6, we will further validate this hypothesis on several personalization tasks derived from individual preferences on real data.

Ensemble reweighting outperforms fine-tuning in the low-data regime. We compare HyRe to model fine-tuning on a synthetic binary classification task. In this task, the training set is generated by first sampling binary labels: label 1 is paired with inputs from $[0, 1]^5$ (all positive inputs), and label 0 is paired with inputs from $[-1, 0]^5$ (all negative inputs). The target distribution is uniform

over $[-1, 1]^5$ with a random linear decision boundary. We compare the performance of ensemble reweighting and fine-tuning by adding the adaptation data to the training set.

Results in Figure 3 show that HyRe consistently outperforms fine-tuning in the low-data regime, achieving high test accuracy with only a few queries. As expected, fine-tuning eventually surpasses ensemble reweighting as the amount of adaptation data increases, due to its higher model capacity. We can view this as a bias-variance tradeoff: reweighting reduces variance while introducing bias by restricting the solution space to functions in the span of the ensemble members. In the low-data regime, the implicit regularization offered by HyRe gives it a clear advantage. Additionally, aside from performance benefits, reweighting is significantly more computationally efficient than fine-tuning, making it especially suitable for large models or resource-constrained settings. The computational cost of HyRe is a single forward pass to compute predictions for each ensemble member, and the cost of computing the ensemble weights (2) is negligible.

5 Related Work

Ensembles and diversity. Many prior works have noted the benefits of ensembles in performance and uncertainty representation, particularly when different ensemble members make different independent mistakes (Krogh & Vedelsby, 1994; Lakshminarayanan et al., 2017). Such findings motivated Mixture-of-Experts (MoE) models which use a gating mechanism to combine predictions from different experts (Jacobs et al., 1991; Jordan & Jacobs, 1994; Yuksel et al., 2012). Recent advancements have incorporated MoE layers into large-scale models, offering computational benefits by conditionally activating a subset of experts (Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2022; Jiang et al., 2024). However, the focus of these works is different from ours, as we expect a small number of experts to outperform the others during evaluation. We leverage recent advances in efficient ensemble training methods (Osband et al., 2023) to train an ensemble of diverse models and propose a method for fast test-time adaptation by working in the space of ensemble weights.

Task (under-)specification and scalable alignment. Task specification is a fundamental aspect of machine learning. While statistical learning theory suggests that expert labels can fully define supervised learning tasks given infinite data (Vapnik, 1999), practical constraints such as limited data and out-of-distribution inputs often lead to task underspecification (Geirhos et al., 2020; D'Amour et al., 2022). Similarly, fully specifying a reward function in reinforcement learning is challenging outside of controlled environments such as games. Overoptimizing for poorly specified rewards can lead to unintended consequences (Zhuang & Hadfield-Menell, 2020; Pan et al., 2022; Skalse et al., 2022; Gao et al., 2023). Instead of specifying a reward function upfront, we can provide human demonstrations, framing task specification as a cooperative game between humans and agents (Hadfield-Menell et al., 2016). This paper introduces a new scalable mode of test-time task specification, which leverages ensemble disagreements to proactively acquire information to resolve ambiguity.

The standard reinforcement learning from human feedback (RLHF) workflow for aligning LLMs (Christiano et al., 2017; Wirth et al., 2017; Ouyang et al., 2022; Rafailov et al., 2024) can be understood as a cooperative game between a human and an agent, where the agent's goal is to learn the human's preferences. Recent works have explored ways to improve the efficiency of RLHF by leveraging the model's uncertainty of human intent for active learning (Ji et al., 2024; Muldrew et al., 2024) and exploration (Dwaracherla et al., 2024). Another line of work on personalization methods (Jang et al., 2023; Li et al., 2024; Poddar et al., 2024) show promise but require per-user preference data, making it necessary to pre-identify user types and collect specific data accordingly. We frame personalization as a special case of task underspecification, demonstrating that a diverse ensemble trained on aggregated data can capture ambiguity, which we can use to directly adapt to new users.

6 Experiments

We conduct several experiments to evaluate the effectiveness of our diverse ensemble approach in various settings, including regression tasks, natural distribution shifts, and personalization scenarios. We defer detailed experimental details to the appendix.

6.1 Regression Data with Distribution Shifts

We evaluate HyRe on three regression datasets from the UCI Machine Learning Repository (Kelly et al.). Specifically, we use the Energy Efficiency, Kin8nm, and CCPP datasets. Building on the

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

Table 1: Root Mean squared error (RMSE) on test data with distribution shifts across three UCI datasets. We compare the performance various ensemble architectures with test-time adaptation using HyRe. We find that HyRe consistently improves the performance of all model architectures.

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 2: Test set accuracy on the Camelyon17 dataset. The DL column indicates whether the algorithm uses domain labels. We see that HyRe engenders competitive performance without use of domain labels.

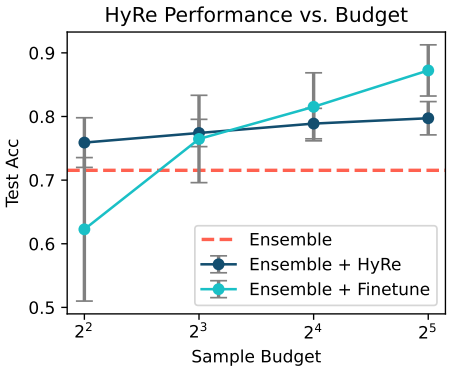

Figure 5: Comparison of HyRe and few-shot fine-tuning on the Camelyon17 OOD test set. HyRe outperforms fine-tuning in the low-data regime despite requiring significantly less computational cost.

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

Table 3: Helpfulness vs Harmlessness tradeoff. The fine-tuned models show an upper bound for performance in each distribution but at the cost of significantly overfitting to one of the two desiderata. HyRe strikes a better balance, improving performance on both metrics.

approach to OOD test set construction in (Sharma et al., 2023), we simulate distribution shift by sorting the data by the average of all input features and assigning samples in the top and bottom 5% of the distribution as a held-out set of out-of-distribution (OOD) samples. The middle 90% of the data is randomly split into a training and validation set. In addition to the two architectures described in Section 2.2, we also evaluate a Vanilla Ensemble model, i.e., a set of independently trained models. Please refer to Appendix C for a detailed description of the Shared Base Ensemble and Epinet architectures. All experiments use ensembles of 100 models, with each architecture employing two MLP layers with 50 units as the prior, base, and learnable components. As an additional point of comparison, we compare HyRe to Monte Carlo Dropout (Gal & Ghahramani, 2016), a representative method for uncertainty estimation. 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 Natural Distribution Shifts

We evaluate a trained Shared-Base ensemble, both with and without HyRe on the WILDS-Camelyon17 dataset (Koh et al., 2021), comparing against several representative methods for OOD generalization from the official WILDS benchmark. As shown in Table 2, test-time adaptation with HyRe consistently outperforms other methods that do not use domain labels and remains competitive with LISA (Yao et al., 2022), 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 5. 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.

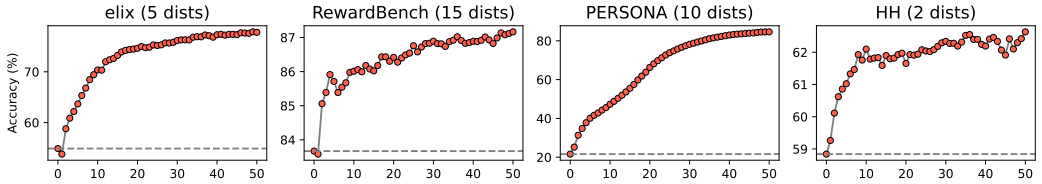

Figure 6: Average reward model accuracy across 18 target distributions from 3 dataset collections. For each collection of preference datasets, we compare the average accuracy of HyRe (red) with different numbers of adaptation samples to the state-of-the-art 2B reward model (dashed line). HyRe consistently outperforms the static reward model with as little as 1-5 labeled examples per distribution.

Model	Type	Overall	Chat	Chat Hard	Safety	Reasoning
Mixtral-8x7B-Instruct-v0.1	DPO	77.6	95.0	64.0	72.6	78.7
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
LLaMA-3-Tulu-2-DPO-70B	DPO	77.2	96.4	57.5	74.9	80.2
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
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)	Seq + Oracle	90.0	92.3	81.8	92.5	93.1

Table 4: **Comparison with state-of-the-art reward models on RewardBench.** Models are categorized by type: DPO (Direct Preference Optimization), Gen (Generative), and Seq (Sequence Classifier). HyRe improves performance over the base GRM-Gemma-2B model with as little as 1-5 labeled examples per distribution.

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 5 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. It is important to note that this fine-tuning serves only as a reference point; our work focuses on test-time adaptation settings where running gradient steps on the model is not feasible.

6.3 Personalizing Preference Models

To evaluate HyRe in personalizing preference models, we evaluate the performance of a 2B reward model on a total of 18 evaluation datasets from three collections of preference data.

Elix. The Elix dataset (Anonymous, 2024), inspired by the "Explain like I'm 5" subreddit, contains 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.

RewardBench. RewardBench (Lambert et al., 2024) is a collection of preference datasets designed to evaluate reward models across various domains, including chat quality, safety, reasoning, coding, and refusal tasks. We use 11 of the provided preference test sets: alpacaeval, donotanswer, hep-go, hep-python, hep-rust, llmbar-adversarial, math-prm, refusals-dangerous, refusals-offensive, xstest-should-refuse, xstest-should-respond.

PERSONA. PERSONA (Castricato et al., 2024) is a dataset for evaluating pluralistic alignment in language models, containing preference data from many synthetic personas with diverse demographic attributes and value systems. We subsample 10 personas from the original dataset for evaluation, treating each persona as a target distribution. We describe dataset details in Appendix F.

Anthropic HH. The Anthropic HH dataset (Bai et al., 2022) contains human preference data used to train and evaluate AI models for helpfulness and harmlessness. We use the helpfulness-base and harmlessness-base splits as evaluation distributions.

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 (Cui et al., 2023) dataset, a standard dataset for reward model training. The base model, a fine-tuned version of Gemma-2B (Team et al., 2024), achieves state-of-the-art accuracy on RewardBench for models at the 2 billion parameter scale, even outperforming GPT-4o (Achiam et al., 2023)[1].

Method	Accuracy
Single Model	0.5903
Conf. Weighted (Jimenez, 1998)	0.6832
Entropy Weighted	0.6838
Logit Ensemble (Jimenez, 1998)	0.8344
Prob Ensemble	0.8365
Majority Vote	0.8371
Shahhosseini et al. (2022) (N=40)	0.8449
Ensemble + HyRe (N=1)	0.8388
Ensemble + HyRe (N=5)	0.8573
Ensemble + HyRe (N=10)	0.8626
Ensemble + HyRe (N=20)	0.8711
Ensemble + HyRe (N=40)	**0.8774**

Figure 7: Comparison of ensemble methods on RewardBench. N indicates number of adaptation samples.

We assess the reward model accuracy when using HyRe to adapt the ensemble model at test time for each of the 18 evaluation datasets, comparing it to the performance of the original reward model. As shown in Figure 6 the initial uniform ensemble performs slightly worse than the original reward model, showing that the ensemble alone is insufficient for performing well across different target distributions. However, the ensemble reweighted with HyRe rapidly adapts to new distributions, surpassing the original reward model's performance with as few as five examples from each distribution. Table 4 shows a detailed comparison of the performance of HyRe against state-of-the-art reward models on the RewardBench leaderboard. Additionally, Figure 7 compares HyRe with several ensemble reweighting methods, illustrating the sample efficiency of HyRe in comparison to existing methods. This experiment demonstrates that HyRe is capable of rapid test-time personalization of large reward models. We show further dataset-level results in the appendix (Figure 9).

We also provide a detailed comparison against models fine-tuned for specific distributions. As an upper bound for performance from targeted fine-tuning, we compare our results with those of models fine-tuned on the helpful-base and harmless-base training sets in the Anthropic-HH dataset. Results in Table 3 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.

7 Discussion

This paper demonstrates that a diverse ensemble can rapidly adapt to new distributions, offering a novel approach to test-time task specification. We think the design space of ensemble architectures for test-time task specification, including mixture-of-experts architectures (Fedus et al., 2022), is a promising direction for future work. Our results in reward modeling show that ensembles can efficiently resolve ambiguities in preferences; future work can close the loop to produce behaviors consistent with new preferences, for example by leveraging the parameterization of Rafailov et al. (2024) which shows a 1-1 correspondence between a reward model and a language model.

Reproducibility Statement. This work uses publicly available pretrained models and datasets. We describe experimental details in Section 6 and Appendix B. To facilitate reproducibility, we will publicly release the data preprocessing pipeline, training code, and experiment scripts alongside the final version of the paper. The public repository will also document all hyperparameters and experimental configurations.

[1] https://huggingface.co/Ray2333/GRM-Gemma-2B-rewardmodel-ft

REFERENCES

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 10]

Anonymous. Elix: Explain like i'm x - a dataset for personalized explanations. 2024. [page 9]

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 10]

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. URL https://doi.org/10.1111/rssb.12158. [page 2, 4, 5]

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018. [page 15]

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 9, 20]

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]

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 10]

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]

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

Vikranth Dwaracherla, Seyed Mohammad Asghari, Botao Hao, and Benjamin Van Roy. Efficient exploration for llms. *arXiv preprint arXiv:2402.00396*, 2024. [page 7]

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 7, 10]

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 8]

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 15]

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]

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

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]

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]

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

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011. [page 15, 19, 20]

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]

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]

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. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/1ab60b5e8bd4eac8a7537abb5936aadc-Paper.pdf. [page 5]

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. [page 7]

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]

Kaixuan Ji, Jiafan He, and Quanquan Gu. Reinforcement learning from human feedback with active queries. *arXiv preprint arXiv:2402.09401*, 2024. [page 7]

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 7]

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 10]

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994. [page 7]

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

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]

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 8, 15]

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. URL https://proceedings.neurips.cc/paper_files/paper/1994/file/b8c37e33defde51cf91e1e03e51657da-Paper.pdf. [page 3, 7]

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, 7, 19]

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 9]

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 15]

Yoonho Lee, Huaxiu Yao, and Chelsea Finn. Diversify and disambiguate: Learning from underspecified data. *International Conference on Learning Representations*, 2023. [page 3, 18]

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 7]

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

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]

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. Active preference learning for large language models. *arXiv preprint arXiv:2402.08114*, 2024. [page 7]

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, 7, 18]

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]

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. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf. [page 3]

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]

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]

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]

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 10]

Mrinank Sharma, Sebastian Farquhar, Eric Nalisnick, and Tom Rainforth. Do bayesian neural networks need to be fully stochastic?, 2023. URL https://arxiv.org/abs/2211.06291. [page 8]

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 7]

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]

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]

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. [page 10]

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, 18]

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 20]

V.N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10 (5):988–999, 1999. doi: 10.1109/72.788640. [page 7]

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 16]

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]

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]

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 8]

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]

Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012. [page 7]

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. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/b607ba543ad05417b8507ee86c54fcb7-Paper.pdf. [page 7]

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 5: 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.

A 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 (Houlsby et al., 2011; Gal et al., 2017).
- **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 (Burda et al., 2018; Laskin et al., 2021). 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.

B 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 (Koh et al., 2021), including the choice of backbone, learning rate, and weight decay.

LLM Preference Learning For our main experiment, we fine-tune the gemma-2b model, specifically the Ray2333/GRM-Gemma-2B-rewardmodel-ft checkpoint. Our fine-tuning experiment uses

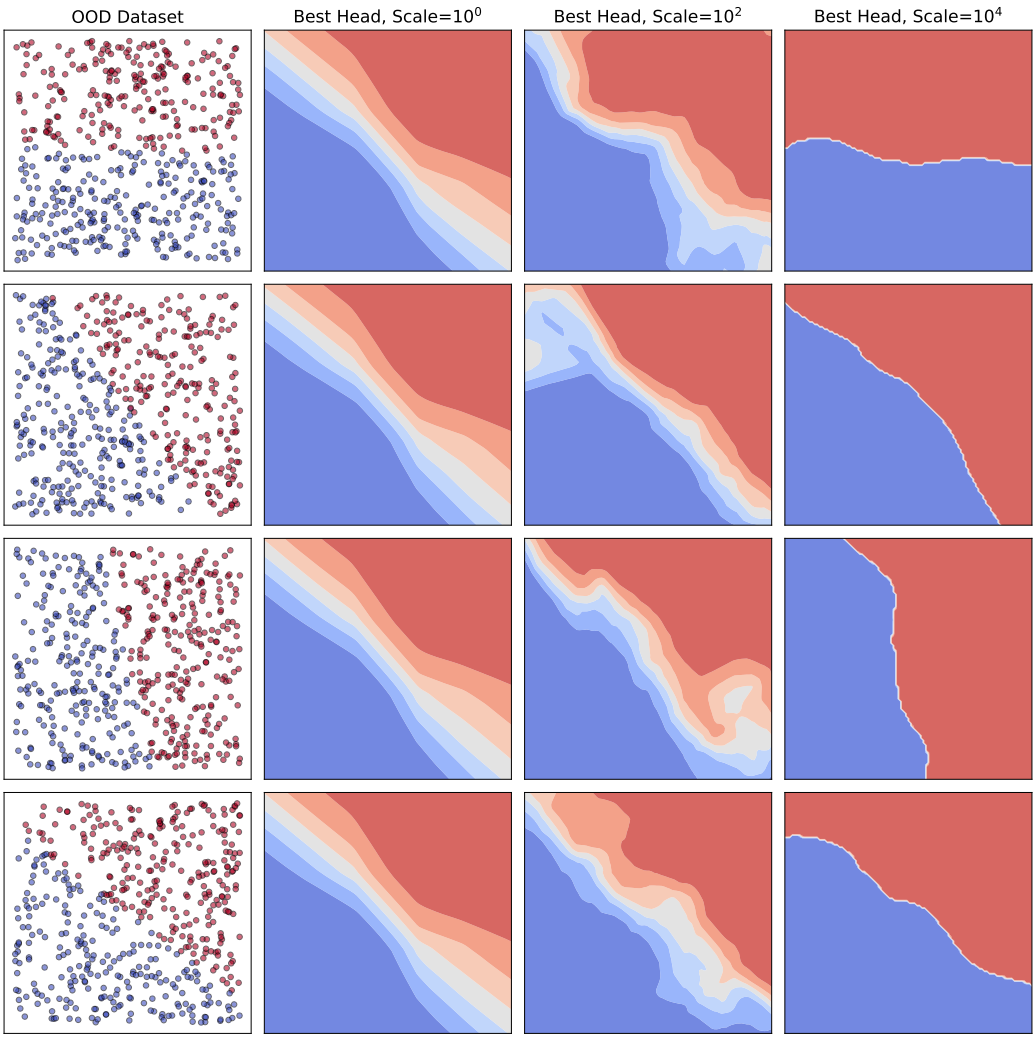

Figure 8: Additional visualizations for the toy conflicting classification example. Increasing the scale hyperparameter results produces heads with sharper decision boundaries.

the base gemma-2b model. We use the TRL codebase for reward model training (von Werra et al., 2020), and use bfloat16 mixed precision for training. We use a learning rate of 0.0001, no weight decay, a batch size of 16, and train for 5000 steps. Our ensemble architecture uses the gemma-2b backbone.

C 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:

$$\sum_{h=1}^{H} \mathcal{L}\left(f_h(x), y\right). \tag{5}$$

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

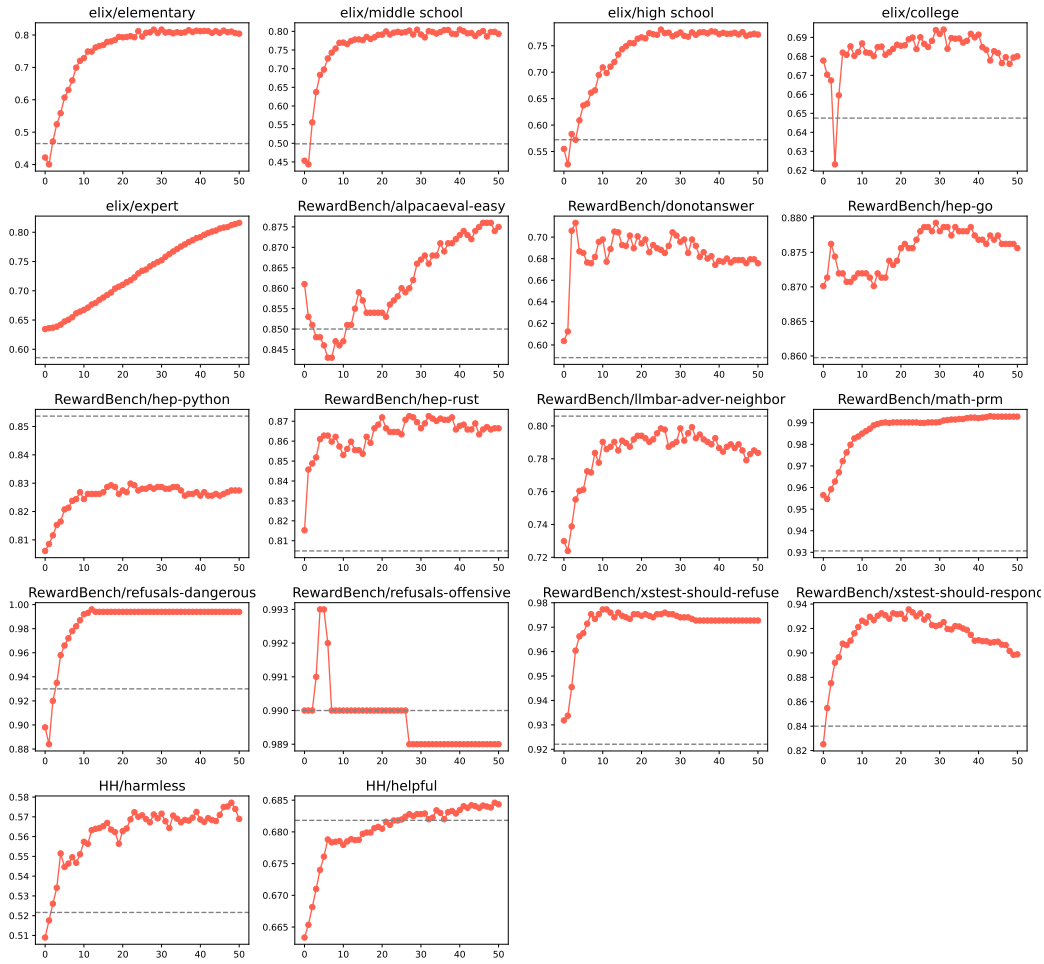

Figure 9: Detailed results for the personalizing preference reward models experiment in Figure 6. Target dataset accuracy (y-axis) after observing different numbers of adaptation samples (x-axis). The dashed line represents the performance of the pretrained reward model.

unseen data in diverse ways. The specific ensemble parameterizations, which we describe below, are designed to achieve this goal.

C.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.

C.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

$$f^p(z) + f^\theta(z) = \begin{bmatrix} f_1^p(z) + f_1^\theta(z) \\ f_2^p(z) + f_2^\theta(z) \\ \vdots \\ f_H^p(z) + f_H^\theta(z) \end{bmatrix} \in \mathbb{R}^H \tag{6}$$

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.

C.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 $f^{\text{epi-frozen}} : \mathcal{X} \to \mathbb{R}^{1,\ldots,d_{\text{index}} \times K}$. 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:

$$f(z, x) = f^{\text{base}}(x) + v f^{\text{epi-frozen}}(x) \cdot z + f^{\text{epi-trainable}}(z, \phi(x), x) \cdot z \tag{7}$$

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., $f_1(x), \ldots, f_H(x) = f(z_1, x), \ldots, f(z_H, x)$. This architecture efficiently generates diverse predictions by sampling different epistemic indices z while leveraging a potentially large pretrained base model.

D 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. (2022) uses a repulsion term that maximizes the cosine distance between the gradient of each ensemble member, and Lee et al. (2023) 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 (Osband et al., 2023) 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.

E 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.

E.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

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

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 (8), 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 (Lakshminarayanan et al., 2017).

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 (Houlsby et al., 2011), 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 (8) 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 $\delta_{nh} = p_{nh} - \frac{1}{H}\sum_h p_{nh}$:

$$\begin{pmatrix} \delta_{11} & \delta_{12} & \delta_{13} & \delta_{14} & \delta_{15} \\ \delta_{21} & \delta_{22} & \delta_{23} & \delta_{24} & \delta_{25} \\ \delta_{31} & \delta_{32} & \delta_{33} & \delta_{34} & \delta_{35} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & -1 & 1 & -1 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \end{pmatrix}. \tag{9}$$

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.

E.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

$$p(x) = \bar{p}(x) + \begin{pmatrix} w_1 & \cdots & w_P \end{pmatrix} \begin{pmatrix} c_{11} & \cdots & c_{1H} \\ c_{21} & \cdots & c_{2H} \\ \vdots & \ddots & \vdots \\ c_{P1} & \cdots & c_{PH} \end{pmatrix} \begin{pmatrix} p_1(x) - \bar{p}(x) \\ p_2(x) - \bar{p}(x) \\ \vdots \\ p_H(x) - \bar{p}(x) \end{pmatrix}, \tag{10}$$

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 (Houlsby et al., 2011).

Factor model. The principal components are maximum likelihood parameters under a linear Gaussian factor model of the data (Tipping & Bishop, 1999). 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" (10), 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.

F PERSONA DATASET DETAILS

Below, we list the personas used in our PERSONA (Castricato et al., 2024) 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.