INTERPRETABILITY IS NOT JUST A TAX: SPARSITY AND DIVERSITY YIELD LEGIBLE AND COMPOSABLE REWARD EXPERTS FOR PERSONALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference modeling aims to capture heterogeneous human values and increasingly relies on learning latent components to represent this variation without fine-grained supervision. What these components actually represent, however, is rarely examined. We audit Mixture-of-Experts (MoE) reward models along three rungs—semantic coherence, functional correspondence, and composability—in controlled and real-world settings, holding the architecture and training data fixed so that differences isolate the effect of the training objective. Our diagnosis is that the components conventional MoE reward models learn are entangled, which limits not only interpretability but also the personalization such models are built for. The entanglement is not intrinsic to the architecture but a property of the training objective: off-the-shelf sparsity and diversity regularizers already substantially reduce it. The resulting configuration, which we call sparse MoE, recovers all underlying semantic categories without domain supervision in a controlled setting, reaching 84.8% expert purity against 60.8% for conventional MoE; on real-world data its experts admit human-interpretable task- and domain-level descriptions aligned with their functional strengths. Targeted interventions on routing weights produce predictable preference changes, showing that these interpretations are behaviorally grounded rather than merely descriptive. Crucially, the disentanglement induced by this objective also makes the experts composable: lightweight router updates yield a 25.81-point improvement in few-shot personalization with 50 examples, roughly 2.7 times the gain of the identical architecture and adaptation procedure applied to entangled experts. Together, these results show that MoE reward models can learn interpretable expert structure, and that the same structure can substantially improve personalization.

1 INTRODUCTION

Reinforcement learning from human feedback (RLHF) commonly relies on reward models to capture human preferences and align large language models (LLMs) with desired behavior (Ziegler et al., 2019; Bai et al., 2022; OpenAI, 2023; Kaufmann et al., 2025). Most reward models, however, represent preferences with a single global function, despite human preferences being heterogeneous and context dependent (Wu, 2025). While earlier approaches model such heterogeneity using fine-grained annotations of preference dimensions, recent work learns multiple latent preference components directly from binary preference data without such supervision and combines them to represent diverse or individualized preferences (Lee et al., 2024; Luo et al., 2025; Shen et al., 2025). This line of work, however, primarily emphasizes predictive or personalization performance, leaving open whether the learned components correspond to coherent and human-interpretable structure.

To address this gap, we ask whether latent components in preference models can become **human-interpretable and functionally specialized**, rather than merely providing additional representational capacity. MoE reward models are a natural setting for this question: their routers expose *where* individual experts are activated, while the experts themselves can be evaluated to determine *what* they are good at (Olson et al., 2025). Across controlled and real-world experiments, we find that entanglement is the default and that it is costly: conventional MoE reward models route almost uniformly and learn experts that behave almost identically, whereas sparsity and diversity regulariz-

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

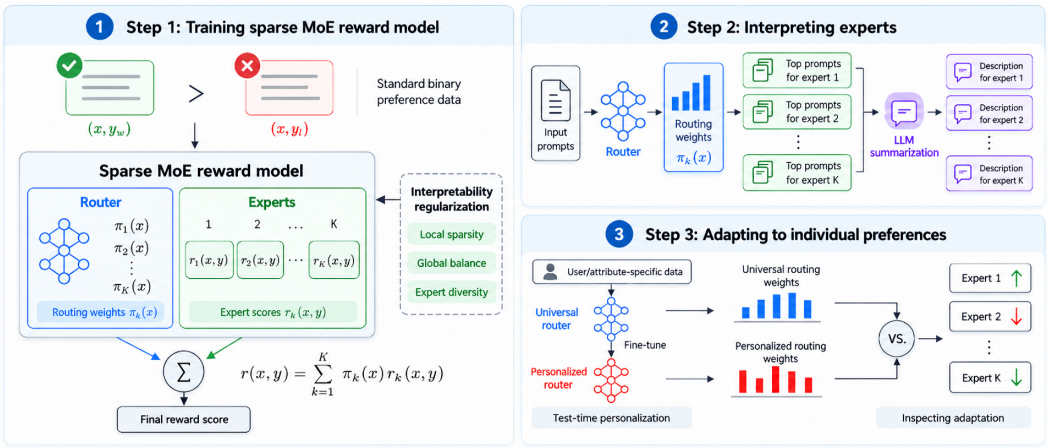

Figure 1: **The models we study and how we analyze them: training an MoE reward model, interpreting its experts, adapting it to individual preferences, and inspecting the adaptation.** An MoE reward model is trained on standard binary preference data with interpretability regularization, and its experts can be interpreted as the domains in which they are activated and specialize. At test time, the router can be fine-tuned to fit personalized preferences, and changes in routing weights show how the model adapts.

ers induce experts whose activation patterns reveal coherent, prompt-identifiable domains and tasks and whose functional strengths align with these interpretations. To establish this, we evaluate latent preference components along three rungs of increasing strength: whether routing patterns are semantically *coherent*, whether they *correspond* to expert strength and behavior, and whether the disentangled components can be *composed* into new preference functions through router updates. We hold the architecture of Shen et al. (2025), the backbone, and the training data fixed and vary only the objective, isolating it as the cause of the structure we observe. Figure 1 summarizes the pipeline: discovering, validating, and composing expert structure.

Our contributions are threefold. **First**, we introduce a three-rung protocol for auditing latent preference components through coherence, correspondence, and composability. **Second**, we show that conventional MoE reward models learn highly entangled experts, while sparsity and diversity regularization produces coherent, functionally specialized experts: in a controlled setting, expert purity rises from 60.8% to 84.8% and specialization from -0.43 to 5.56, with similar gains also on real-world preference data. Targeted interventions further confirm that these interpretations reflect genuine expert behavior. **Third**, we show that this specialization also improves personalization: with 50 adaptation examples, sparse MoE gains 25.81 points on RPR, compared with 9.57 points for the same architecture without regularization. Thus, the same specialization that makes experts interpretable also makes them more composable for personalization.

2 Related Work

Latent Preference Modeling and Personalization Most existing reward models are trained to reflect universal human values, yet they often fail to capture their heterogeneous nature and are therefore not well-suited for representing individualized preferences (Chakraborty et al., 2024; Zhang et al., 2025). Many recent approaches introduce multiple preference components with additional supervision, such as fine-grained attribute annotations (Wang et al., 2024a; Quan, 2024; Zhou et al., 2024; Wu & Lu, 2026) or user-specific histories (Zhao et al., 2024; Choi et al., 2025; Liu et al., 2025; Chen et al., 2025a), and represent unseen preferences at test time by reweighting these predefined attributes or prototypes. While effective, they depend on costly extra annotation.

To reduce this requirement, recent work learns latent preference structure directly from binary preference data. Some approaches introduce latent user variables, via variational autoencoders (Poddar et al., 2024) or expectation-maximization over user prototypes (Chidambaram et al., 2026). Others learn multiple reward components and adapt their mixture, through ensembles of reward models (Lee et al., 2024) or principal component analysis of response representations (Luo et al., 2025).

Closest to our work is MiCRo (Shen et al., 2025), which trains an MoE reward model on standard preference data and adapts its router to context-dependent preferences. These approaches show that latent components improve preference modeling, but leave open what the components represent: whether they are coherent, human-interpretable, and aligned with distinct functional roles.

Interpretability through Sparsity and MoE Prior work has shown that sparsity can help disentangle LLM representations and recover more interpretable units (Elhage et al., 2022; Bricken et al., 2023). Sparse autoencoders (SAEs), for example, project representations into a higher-dimensional space under sparsity constraints, yielding monosemantic features that are interpreted by summarizing highly activating examples (Shu et al., 2025), and these features have been used to build interpretable reward functions (Movva et al., 2026; Liu et al., 2026; Zhang et al., 2026). Such pipelines learn features separately from the reward objective, so the features are not necessarily preference-relevant, and their weights are typically fixed across inputs, limiting flexible preference modeling.

In parallel, MoE architectures have emerged as an effective paradigm for scaling LLMs (Cai et al., 2025). Their explicit routing structure also provides a natural lens for interpretability, as routing patterns often align with semantic structures, and experts exhibit emergent specialization (Fayyaz et al., 2026). MoE reward models are therefore a natural setting for studying latent preference components: unlike SAE-based pipelines, their expert structure is learned jointly with the reward objective, and their context-dependent routing allows experts to be recombined across inputs. Yet beyond Wang et al. (2024b) and Shen et al. (2025) they remain largely underexplored, and scalar preference supervision provides no explicit semantic signal for how latent experts should organize.

Following Do et al. (2026), we interpret experts as the domains and contexts in which they are preferentially activated and exhibit specialization, such as math problem solving or feedback seeking, and ask whether such structure can emerge from binary preference supervision alone. This notion of interpretation identifies *where* an expert is activated and functionally specialized. Such structure is naturally useful for personalization: if experts specialize in different contexts, reweighting and recombining them is a natural way to emphasize different behaviors for different users. It also enables targeted control and auditing of how reward models organize and adapt their preferences.

3 MoE Reward Models and Interpretability Regularization

Preliminary In RLHF, a reward model $r_\theta(x, y) \in \mathbb{R}$ estimates human preferences over responses y given a prompt x, and is trained on pairwise comparisons $\mathcal{D} = \{(x, y_w, y_l)\}$ in which y_w is preferred over y_l. It is typically a Bradley–Terry model (Bradley & Terry, 1952), which defines the preference probability as:

$$P(y_w \succ y_l \mid x) = \sigma\big(r_\theta(x, y_w) - r_\theta(x, y_l)\big), \tag{1}$$

where $\sigma(z) = (1 + e^{-z})^{-1}$. The model is trained by minimizing the negative log-likelihood of observed preferences.

MoE Reward Model Wang et al. (2024b) and Shen et al. (2025) propose using MoE architectures to model diverse and heterogeneous preferences. Specifically, an MoE reward model consists of K experts $\{r_{\theta_k}\}_{k=1}^K$ and a router $\pi_\phi(x) \in \Delta^{K-1}$. In practice, we implement experts as linear heads and the router as a one-hidden-layer MLP over backbone model representations. The router produces a softmax distribution over experts based on the input x, where $\pi_{\phi,k}(x)$ denotes the weight assigned to expert k. The final reward is computed as the weighted average of expert scores:

$$r(x, y) = \sum_{k=1}^{K} \pi_{\phi,k}(x) \cdot r_{\theta_k}(x, y). \tag{2}$$

During training, we optimize the marginalized expert-level BT loss following Shen et al. (2025):

$$\mathcal{L}_{\mathrm{RM}} = -\frac{1}{|\mathcal{D}|} \sum_{(x,y_w,y_l)\in\mathcal{D}} \log \sum_{k=1}^{K} \big(\pi_{\phi,k}(x) \cdot \sigma(r_{\theta_k}(x, y_w) - r_{\theta_k}(x, y_l))\big). \tag{3}$$

Compared to applying BT loss to the expected expert score, the marginalized expert-level loss assigns comparison-level gradients to individual experts rather than broadcasting the same preference signal to all active experts, encouraging more diverse expert behaviors.

Interpretability Regularization To test whether the entanglement of conventional MoE stems from the architecture or from the training objective, we hold the architecture fixed and add three standard interpretability regularizers to the objective, encouraging disentangled routing and the emergence of expert specialization; we refer to this configuration as sparse MoE. Using established regularizers keeps the manipulation minimal and interpretable: structure that emerges can be attributed to the objective itself rather than to the particular mechanism that implements it.

(1) Local Sparsity: Following prior work on interpretability through sparsity (Elhage et al., 2022; Bricken et al., 2023), we encourage sparse routing distributions by minimizing the entropy of the routing weights:

$$\mathcal{L}_{\text{ls}} = \mathbb{E}_{x\sim\mathcal{D}}\left[\frac{H\big(\pi_\phi(x)\big)}{\log K}\right], H(\pi) = -\sum_{k=1}^{K}\pi_k \log \pi_k. \tag{4}$$

(2) Global Balance: To prevent model collapse, we encourage balanced utilization across experts by maximizing the entropy of the average routing distribution within a batch $\mathcal{B}$:

$$\mathcal{L}_{\text{gb}} = \mathbb{E}_{\mathcal{B}\sim\mathcal{D}}\left[-\frac{H(\bar{\pi}_\phi)}{\log K}\right], \bar{\pi}_\phi = \mathbb{E}_{x\sim\mathcal{B}}[\pi_\phi(x)]. \tag{5}$$

(3) Expert Diversity: We encourage experts to capture distinct preference signals by penalizing correlations between their reward differences. Let $\Delta_k = r_{\theta_k}(x, y_w) - r_{\theta_k}(x, y_l)$, and we minimize the average squared pairwise Pearson correlation within a batch $\mathcal{B}$ of preference pairs:

$$\mathcal{L}_{\text{div}} = \frac{2}{K(K-1)}\sum_{i<j}\text{corr}(\Delta_i, \Delta_j)^2. \tag{6}$$

The overall training objective is defined as $\mathcal{L} = \mathcal{L}_{\text{RM}} + \lambda_{\text{ls}}\mathcal{L}_{\text{ls}} + \lambda_{\text{gb}}\mathcal{L}_{\text{gb}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}}$, where λ_{ls}, λ_{gb}, and λ_{div} control the strengths of local sparsity, global balance, and expert diversity.

4 Evaluation Protocol: Coherence, Correspondence, and Composability

We evaluate whether latent preference components are interpretable along three rungs, each answering a question the previous one leaves open. *Coherence* asks whether routing patterns admit a consistent semantic description, measured by category recovery and expert purity in the controlled setting (Section 5.1) and by description fidelity and human ratings on real-world data (Section 6.1). *Correspondence* asks whether those descriptions track expert behavior, tested at two levels: *correlationally*, by an *expert specialization* score measuring whether an expert outperforms its peers on the inputs routed to it (Sections 5.1 and 6.1), and *causally*, by targeted attribute steering (Section 5.2). *Composability* asks whether disentangled components can be combined into new preference functions, measured by test-time preference adaptation (Section 6.2). Each rung rules out a weaker reading of the one below: coherence alone may reflect only a partition of the inputs; a correlational association may reflect a router that tracks input statistics rather than experts with distinct roles; and control over individual experts may still not extend to combining them for preferences the model was never trained on.

Models Following Shen et al. (2025), we use a strong reward model GRM-Llama3.2-3B (Yang et al., 2024) as the frozen backbone and train K linear heads as experts. Unless otherwise specified, we set λ_{ls}, λ_{gb}, and λ_{div} to 0.5, 0.5, and 2.0. The number of experts is set per experiment according to the data complexity. Appendix A details the setups; Appendices B and C ablate the expert number K and the regularization terms and give guidelines for selecting hyperparameters; Appendix D repeats the real-world experiments with Skywork-Reward-V2-Qwen3-8B as the backbone.

Baselines A single-head reward model serves as a reference for universal preference modeling, and among MoE models we compare vanilla MoE, MiCRo (Shen et al., 2025), which encourages a uniform routing distribution, and sparse MoE. In the real-world setting we add two SAE-based baselines, SARM (Liu et al., 2026) and WIMHF (Movva et al., 2026); because their interpretation mechanisms differ, we exclude them from the human evaluation and report personalization only for SARM. Appendix E compares MoE- and SAE-based reward models directly.

5 DIAGNOSING EXPERT STRUCTURE: CONTROLLED SETTINGS

We first consider two controlled settings in which the underlying structure is known, which lets us test the first two rungs of the protocol (Section 4) directly: whether the router recovers semantically coherent structure, whether the associated experts are strongest where they are routed, and whether intervening on them changes preferences as predicted.

5.1 RECOVERING LATENT SEMANTIC STRUCTURE

We investigate whether MoE can discover latent semantic structure in the data without domain supervision. We use the Stanford Human Preferences (SHP; Ethayarajh et al., 2022) dataset, which includes annotations of the prompt domains. As these domains vary in granularity, we group the 16 subsets into four categories: *academic*, *professional*, *cooking*, and *fiction*.

We train MoE reward models with $K = 5$ experts on the dataset without using the domain annotations. After training, we assign each expert a category label according to the majority category among its 50 highest-routed validation examples. We then evaluate (i) whether routing patterns recover the held-out category structure and (ii) whether experts are functionally specialized in the categories to which they are preferentially routed.

Evaluation Setting In addition to the task accuracy, we report two model behavior metrics: average routing entropy (Eq. 4) and average expert correlation (reported without squaring; Eq. 6 uses its squared form for training). Values closer to zero indicate sparser routing and more distinct expert behavior, respectively, which are desirable for routing-pattern interpretability and expert specialization. Since the latent categories are known, we further evaluate interpretability and specialization directly using the following metrics adapted from prior work (Badjie et al., 2025; Yang et al., 2025; Herbst et al., 2026):

- *Category Coverage*: the proportion of categories assigned to at least one expert, measuring how much of the underlying structure the router captures.
- *Expert Purity*: for each expert, the proportion of its 50 top-activating test examples that belong to its assigned category; higher values mean more coherent routing.
- *Expert Specialization*: for each expert, the gap between its accuracy on its labeled category and the average accuracy of all experts on that category; higher values mean the expert is functionally stronger where it is routed.

Table 1: **Latent structure recovery and expert specialization on SHP.** Without access to domain labels during training, sparse MoE recovers coherent routing structure and develops experts that are functionally specialized in their routed categories. Results are averaged over three runs.

		Model Behavior		Interpretable Structure		
Method	Accuracy (↑)	Routing Entropy (↓)	Expert Correlation (→ 0)	Category Coverage (↑)	Expert Purity (↑)	Expert Specialization (↑)
Single Reward	70.71	-	-	-	-	-
Vanilla MoE	**72.88**	0.996	0.594	0.67	60.80	-0.43
MiCRo	72.64	1.000	0.525	0.42	50.27	0.24
Sparse MoE	70.18	**0.157**	**0.010**	**1.00**	**84.76**	**5.56**

Results Table 1 shows a clear difference in the structure learned by the three MoE variants. Sparse MoE is the only model that recovers all four latent categories, despite receiving no category supervision during training. By contrast, vanilla MoE and MiCRo exhibit nearly uniform routing and reach category coverage of only 0.67 and 0.42. Sparse MoE's routing is also far more coherent: 84.76% of each expert's highest-routed examples belong to that expert's assigned category, against 60.80% for vanilla MoE and 50.27% for MiCRo. Crucially, this structure comes with distinct expert behavior: sparse MoE's experts beat the average expert by 5.56 points on their assigned categories, whereas vanilla MoE's and MiCRo's match or trail it. The discovered categories are therefore not merely patterns in input assignment; they correspond to experts with different functional strengths. Regarding task performance, sparse MoE achieves 70.18% overall accuracy, compared with 72.88% for vanilla MoE and 72.64% for MiCRo. Thus, the stronger structural organization comes with a modest 2–3 point reduction in aggregate accuracy. The drop is concentrated in the *academic* and *professional* categories, which are broad and heterogeneous: there the strongest expert for an input

often does not receive the highest routing weight (Appendix F). The cost therefore reflects imperfect routing rather than a limit on what the experts themselves can represent.

Table 2: **Input prompts together with the sparse MoE routing weights assigned to each expert.** Parentheses under each expert indicate the category associated with that expert. The router produces sparse and semantically meaningful routing distributions based on the input semantics.

Input Prompt	Category	Expert 0 (Academic)	Expert 1 (Academic)	Expert 2 (Professional)	Expert 3 (Cooking)	Expert 4 (Fiction)
Why is 18 the maximum amount of electrons an atomic shell can hold?	Academic	0.025	**0.967**	0.000	0.007	0.000
Why is it that Eastern and Western philosophy don't appear to have interacted much, even in the ancient world, until the 19th century? Is there any evidence of cross pollination of ideas?	Academic	**0.777**	0.222	0.000	0.000	0.001
I own a condo and my downstairs neighbors have called police on me over 300 times in the last 365 days claiming frivolous noise complaints, HOA is now involved and threatening legal action. Can they force me out of my unit? No mortgage, bought in cash...I am sole occupant. (MN)	Professional	0.000	0.000	**0.989**	0.000	0.011
A friend said I should crack my garlic 10 minutes before cooking, is there a real reason for this? She said that it causes some chemical reaction to make the garlic taste better, can I just cut the garlic up and wait 10 minutes or is this pointless to begin with?	Cooking	0.000	0.123	0.000	**0.877**	0.000
[Marvel Comics] Why did Norse paganism fade away, when Thor was actually present on Earth, fighting monsters and performing miracles?	Fiction	0.178	0.000	0.000	0.000	**0.822**

Qualitative Analysis Table 2 shows representative SHP prompts together with their sparse MoE routing distributions. The router typically assigns most probability mass to a single expert, and the dominant expert is consistent with the semantic category inferred from that expert's highest-routed examples. Routing need not be strictly one-hot: some prompts assign non-negligible mass to a secondary expert when an input involves multiple semantic aspects, suggesting that the model combines several learned components when appropriate.

The learned structure can also be finer-grained than the four evaluation categories. Within the broad *academic* category, the science question is routed primarily to Expert 1, whereas the humanities question is routed primarily to Expert 0. Although both receive the same coarse category label, their routing behavior suggests that the model has discovered finer semantic distinctions than the evaluation labels represent.

Takeaway (coherence and correspondence): Without domain supervision, sparse MoE recovers human-recognizable latent structure from binary preference data, and the resulting routing patterns align with distinct expert strengths.

5.2 Functional Validation through Targeted Attribute Steering

The correspondence established so far is correlational: experts are strongest on the inputs routed to them. We now test it causally, through targeted post-hoc interventions on routing weights in a controlled setting with known preference attributes. If routing patterns identify genuine expert roles, intervening on the corresponding experts should produce predictable and targeted changes in model behavior.

We train MoE models with $K = 10$ experts on the Reasonable Preference Reversal (RPR; Pitis et al., 2024) dataset, which spans 40 attributes and whose relative preference reverses under each (e.g., clarity vs. creativity). Following Shen et al. (2025), we append a specification of the desired attribute to each prompt during training so the model learns attribute-conditioned preferences. The attribute enters only as input context: the objective remains the binary preference loss, no expert is assigned an attribute, and the expert–attribute associations used below are derived post hoc from routing statistics. At test time we remove that specification, so preference changes are induced solely by interventions on the router. For each attribute we select the target expert whose top 100 activating validation examples contain the highest proportion of that attribute—our interpretation of the expert's role—set its weight to $w \in \{0.25, 0.5, 0.75, 1.0\}$, and rescale the remaining weights to keep the distribution normalized. A further baseline selects target experts at random.

Evaluation Setting We evaluate steering effectiveness using two metrics: *desired flip rate* and *undesired flip rate*. For each attribute, we first identify the sets of examples that are correctly and incorrectly classified when the target expert is removed. The desired flip rate is defined as the proportion of initially incorrect examples that switch to the correct response after assigning a specified

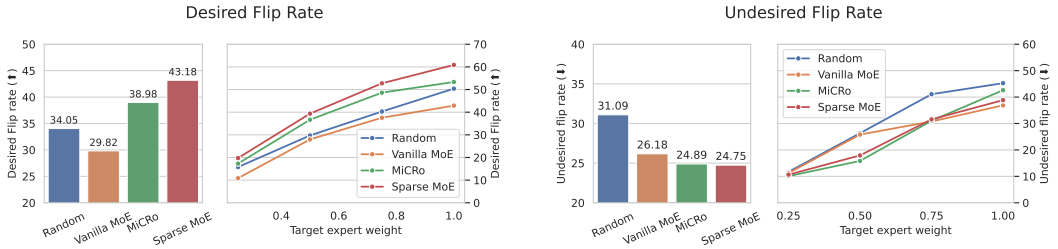

Figure 2: **Attribute steering results on RPR.** Sparse MoE achieves the highest desired flip rate (left) while keeping a low undesired flip rate (right), showing strong and targeted preference steering. Results are averaged over three runs.

weight to the target expert. Conversely, the undesired flip rate measures the proportion of initially correct examples that become incorrect after manipulating the target expert's weight. Intuitively, effective and targeted steering should achieve a high desired flip rate without increasing the undesired flip rate. The final scores are averaged over all target weights and attributes.

Results Figure 2 shows the desired and undesired flip rates both on average and as a function of the target expert weight. Sparse MoE achieves the highest desired flip rate, outperforming the strongest baseline by 4.20 points on average and exceeding 60% when the target expert receives full routing weight ($w = 1.0$). The rate also rises consistently as the target expert is upweighted, so intervening on routing-identified experts produces predictable shifts toward the corresponding attributes. Two effects contribute. First, sparse MoE's experts are behaviorally distinct, which makes routing interventions more influential in general: even randomly selected experts produce substantial preference changes when upweighted. Second, selecting experts by their routing-based attribute associations improves the desired flip rate further, indicating that the semantic patterns revealed by routing carry real information about experts' roles. Sparse MoE also keeps a low undesired flip rate, so its interventions are targeted rather than indiscriminate, whereas vanilla MoE steers substantially more weakly, consistent with its limited specialization, and MiCRo falls in between. The routing patterns learned by sparse MoE are therefore behaviorally grounded: experts identified from their activations can be manipulated to change preferences predictably.

Takeaway (causal correspondence): The semantic structure recovered by sparse MoE is not merely descriptive; routing-based expert interpretations identify functionally distinct components whose targeted intervention predictably changes model preferences.

6 Diagnosing Expert Structure: Real-World Settings

Having established coherence and correspondence where the latent structure is known, we ask whether the same structure emerges in real-world preference data, where no ground-truth decomposition exists. We train and interpret MoE models on preference-700K (700K; Dong et al., 2024), using $K = 20$ experts and $\lambda_{\text{gb}} = 1.0$ for sparse MoE to capture diverse patterns without routing collapse; Appendix D reports the same analyses for LMArena.

6.1 Human-Interpretable and Functionally Specialized Experts

Following prior work (Shu et al., 2025), we interpret each expert by summarizing its top-activating examples: we collect the 20 validation prompts with the largest routing weights and prompt an LLM (Qwen3.6-27B) to describe their shared characteristics in natural language. This description is our routing-based interpretation of the expert; Table 3 shows examples for sparse MoE.

Evaluation Setting Without ground-truth expert labels, we test coherence and correlational correspondence with two metrics. For routing pattern interpretability, we measure *description fidelity* following Movva et al. (2026). For each expert, an LLM (Qwen3.6-35B-A3B) judges whether 200 test examples match its natural-language description, and we compute the Pearson correlation between these judgments and the corresponding routing weights. Higher fidelity indicates that routing patterns are coherent and well captured by the expert description. For functional specialization, we

Table 3: **Natural language descriptions of experts in sparse MoE.** Highest-routed examples reveal topic- and domain-level patterns (all descriptions in Appendix F).

ID	Natural Language Description	Description Fidelity (↑)	Expert Specialization (↑)
0	asks for advice, feedback, or discussion on a personal, professional, or academic challenge	0.43	22.08
6	asks to solve a procedural task involving numerical or list manipulation	0.51	25.98
8	asks for a summary or extraction of information from a provided text	0.45	19.35
10	asks for a direct answer to a multiple-choice or short-answer question	0.36	20.56
13	requests assistance with illegal, harmful, or sexually explicit activities	0.62	3.35
15	asks to solve a competitive mathematics problem	0.42	32.00

adapt the *expert specialization* score from Section 5.1. For each expert, we compare its accuracy on its top 10% highest-routed examples with the average accuracy of all experts on the same examples. A higher score indicates that the expert is functionally stronger where it is preferentially routed.

Results Table 4 shows that sparse MoE learns substantially more coherent and functionally differentiated expert structure than the baselines. Its description fidelity is the highest (0.38), indicating that its routing patterns are more consistently captured by natural-language interpretations. More importantly, these semantic patterns align with distinct expert strengths: sparse MoE reaches an expert specialization score of 15.13, whereas vanilla MoE and MiCRo remain near zero. The learned interpretations therefore describe not only where an expert is activated but also the contexts in which it performs particularly well, and Table 3 shows that these contexts correspond to recognizable tasks and domains. Sparse MoE retains competitive accuracy (83.78%) against vanilla MoE (84.82%) and MiCRo (84.68%), while the SAE baselines are weaker on both axes.

Table 4: **Semantic coherence and functional specialization.** Sparse MoE performs the best in both while preserving strong accuracy.

Method	Accuracy (↑)	Description Fidelity (↑)	Expert Specialization (↑)
Single Reward	83.52	-	-
SARM	75.72	0.09	2.48
WIMHF	60.43	0.21	-9.06
Vanilla MoE	**84.82**	0.29	0.09
MiCRo	84.68	0.15	0.04
Random	-	0.01	-0.04
Sparse MoE	83.78	**0.38**	**15.13**

Human Evaluation We also check whether the routing patterns are coherent to humans. For each model, we randomly sample five experts and present their ten highest-routed examples to five annotators, who rate *pattern coherence* and *description quality* on a 1–5 scale (Appendix G). Sparse MoE receives the highest ratings on both dimensions, with the largest gain in pattern coherence (Figure 3), corroborating the automatic evaluation.

Takeaway (coherence and correspondence at scale): On real-world preference data, sparse MoE discovers interpretable routing structure aligned with distinct expert strengths.

6.2 Composing Interpretable Experts for Preference Adaptation

We now turn to the third rung: whether the structure is composable as well as legible. We conduct test-time preference adaptation, freezing the experts and updating only the router on a small adaptation set.

Evaluation Setting Following Luo et al. (2025), we target the seven RPR attributes with the most examples, using $n = 50$ training examples for adaptation and evaluating on all test examples. We adapt the MoE models trained on 700K, which never see RPR's attribute annotations, freezing all experts and updating only the router with the Hedge algorithm (Shen et al., 2025). We also compare against HyRe (Lee et al., 2024), an ensemble of 20 reward models whose combination weights are adapted at test time, and ArmoRM (Wang et al., 2024b), an 8B MoE reward model whose experts are trained with attribute annotations. We report post-adaptation accuracy and improvement over the corresponding unadapted model, averaged over three independently sampled adaptation sets.

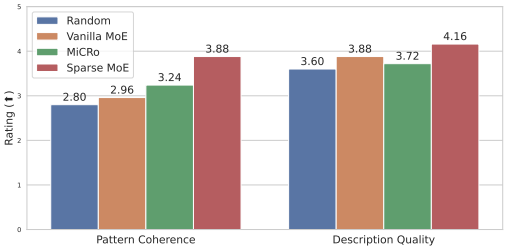

Figure 3: **Human evaluation of model interpretability.** Annotators find sparse MoE has the most interpretable routing patterns. The improvements in pattern coherence are statistically significant ($p < 0.05$).

Results Table 5 shows that sparse MoE benefits substantially more from router adaptation than the competing methods. With only 50 adaptation examples its overall accuracy rises by 25.81 points to 72.10%, against 15.33 for SARM and roughly 9–10 for vanilla MoE and MiCRo, giving it both the largest gain and the highest post-adaptation accuracy. Because vanilla MoE and MiCRo share sparse MoE's router and adaptation algorithm, this gap isolates the effect of expert structure rather than of the adaptation interface; HyRe, which adapts 20 ensemble weights with no expert specialization, gains only 5.98 points. The gains are largest on attributes where the unadapted model performs poorly, indicating that the experts can be recombined to represent preferences far from the universal preference function. Moreover, Appendix I shows that the sparse MoE experts that receive the largest weight increases after adaptation often exhibit plausible semantic associations with the target attributes, further linking effective personalization to expert interpretability.

Table 5: **Personalization results on RPR.** Scores in parentheses denote changes after adaptation. Sparse MoE shows the strongest overall personalization performance. Appendix H further shows that a 14-point improvement is already reached with 5 examples.

Method	Linguistic Creativity	User Friendliness	Humor & Entertainment	Scientific Rigor	Narrative & Storytelling	Creativity & Originality	Factual Accuracy	Overall
Single Reward	31.73	39.33	30.16	76.19	34.17	18.52	85.92	45.15
SARM	62.82 (+19.55)	**60.30** (+13.11)	**83.73** (+50.40)	62.70 (-3.97)	60.83 (+10.83)	**61.11** (+31.94)	57.28 (-14.55)	64.11 (+15.33)
HyRe	37.50 (+6.73)	46.44 (+9.36)	32.14 (+3.57)	78.17 (+0.79)	40.00 (+7.50)	31.02 (+14.35)	81.22 (-0.47)	49.50 (+5.98)
Vanilla MoE	42.31 (+11.54)	44.19 (+8.23)	36.51 (+6.75)	81.74 (+3.17)	51.67 (+19.17)	28.70 (+10.64)	**89.20 (+7.51)**	53.47 (+9.57)
MiCRo	40.39 (+9.62)	42.32 (+6.36)	40.48 (+10.72)	81.35 (+2.78)	51.25 (+18.75)	28.24 (+10.18)	88.73 (+7.04)	53.25 (+9.35)
ArmoRM	21.15 (+0.00)	42.32 (+2.99)	14.29 (+0.00)	**84.52** (+0.00)	32.50 (+0.00)	23.61 (+1.39)	81.69 (+0.00)	42.87 (+0.63)
Sparse MoE	**74.04 (+43.27)**	55.06 **(+14.61)**	80.56 **(+50.80)**	82.54 **(+7.54)**	**68.75 (+28.75)**	60.19 **(+31.02)**	83.57 (+4.70)	**72.10 (+25.81)**

The personalization improvement is also visible at the user level: on PersonalLLM (Zollo et al., 2025), sparse MoE gains 2.47 points after adaptation whereas vanilla MoE and MiCRo *lose* accuracy (-0.74 and -0.90; Appendix J), so entangled components are not only harder to interpret but also harder to recombine. Appendix D reports similar gains across backbones and datasets. The adapted reward models can also serve as optimization signals for policy training: for two attributes, GRPO post-training against an adapted sparse MoE reward model yields responses preferred over those from a policy trained against the unadapted model (53.0 versus 45.5 win rate on creativity, 52.0 versus 44.5 on scientific rigor), without reducing overall quality (Appendix K).

Takeaway (composability): The disentangled, specialized experts sparse MoE learns are also composable: lightweight router adaptation recombines them for unseen preferences. The same adaptation gains far less on entangled experts, and degrades at the user level.

7 Conclusion

We investigated whether latent experts in MoE reward models can become both interpretable and useful for personalization. Conventional training leaves experts entangled, whereas simple sparsity and diversity regularization yields coherent, functionally specialized experts whose roles are supported by targeted interventions. The same specialization also makes experts more composable, leading to substantially stronger few-shot personalization across settings. Although this comes with a modest reduction in universal preference accuracy, the resulting expert structure provides a more effective basis for adapting reward models to unseen preferences. [1]

AI Use Statement

In this work, we used generative AI tools for coding assistance, text polishing, related work discovery, and figure generation. We have not used generative AI tools for idea discovery or development of the conceptual framework. All scientific content, experimental decisions, analyses, and conclusions were developed and verified by the authors.

[1] Limitations are discussed in Appendix L.

Reproducibility Statement

The code, data, configurations, and prompts used to implement and reproduce our experiments are described in the paper and included with the submission. We will make the code base publicly accessible after the paper is accepted.

Ethics Statement

While our work aims to address the challenge of modeling heterogeneous human values, the models we study also introduce potential risks of dual use, as they can be adapted to overfit to individual preferences. As a result, universal principles such as fairness and safety constraints could potentially be bypassed, and harmful behaviors or addictive usage patterns may be reinforced. In addition, making certain preference patterns more interpretable may increase the risk of privacy leakage or misuse of sensitive user information. Therefore, we strongly recommend that any real-world deployment of such systems undergo careful ethical review and oversight by appropriate ethics committees.

References

Bakary Badjie, José Cecílio, and António Casimiro. Double-stage feature-level clustering-based mixture of experts framework. *CoRR*, abs/2503.09504, 2025. doi: 10.48550/ARXIV.2503.09504. URL https://doi.org/10.48550/arXiv.2503.09504.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862, 2022. doi: 10.48550/ARXIV.2204.05862. URL https://doi.org/10.48550/arXiv.2204.05862.

Avinandan Bose, Zhihan Xiong, Yuejie Chi, Simon Shaolei Du, Lin Xiao, and Maryam Fazel. Lore: Personalizing LLMs via low-rank reward modeling. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=bYu4DOqRY8.

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

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*, 37(7):3896–3915, 2025. doi: 10.1109/TKDE.2025.3554028.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. MaxMin-RLHF: Alignment with diverse human preferences. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6116–6135. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chakraborty24b.html.

Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. PAL: Sample-efficient personalized reward modeling for pluralistic alignment. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=1kFDrYCuSu.

Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. PAD: Personalized alignment at decoding-time. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=e7AUJpP8bV.

Keertana Chidambaram, Karthik Vinay Seetharaman, and Vasilis Syrgkanis. Direct preference optimization with unobserved preference heterogeneity: The necessity of ternary preferences. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026. URL https://openreview.net/forum?id=7RFcY3ljMO.

Youngbin Choi, Seunghyuk Cho, Minjong Lee, MoonJeong Park, Yesong Ko, Jungseul Ok, and Dongwoo Kim. CoPL: Collaborative preference learning for personalizing LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12875–12893, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.650. URL https://aclanthology.org/2025.emnlp-main.650/.

Haikang Deng and Colin Raffel. Reward-augmented decoding: Efficient controlled text generation with a unidirectional reward model. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11781–11791, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.721. URL https://aclanthology.org/2023.emnlp-main.721/.

Giang Do, Hung Le, and Truyen Tran. Do domain-specific experts exist in MoE-based LLMs? In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2324–2340, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.110. URL https://aclanthology.org/2026.findings-acl.110/.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. RLHF workflow: From reward modeling to online RLHF. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a13aYUU9eU.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 17–23 Jul 2022.

Mohsen Fayyaz, Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Ryan A. Rossi, Trung Bui, Hinrich Schuetze, and Nanyun Peng. Steering moe LLMs via expert (de)activation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=v5Yl9V8rJs.

Jeremy Herbst, Jae Hee Lee, and Stefan Wermter. The expert strikes back: Interpreting mixture-of-experts language models at expert level. *CoRR*, abs/2604.02178, 2026. doi: 10.48550/ARXIV.2604.02178. URL https://doi.org/10.48550/arXiv.2604.02178.

James Y. Huang, Sailik Sengupta, Daniele Bonadiman, Yi-An Lai, Arshit Gupta, Nikolaos Pappas, Saab Mansour, Katrin Kirchhoff, and Dan Roth. DeAL: Decoding-time alignment for large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26280–26300, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1274. URL https://aclanthology.org/2025.acl-long.1274/.

Yuki Ichihara, Yuu Jinnai, Tetsuro Morimura, Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Eiji Uchibe. Evaluation of best-of-n sampling strategies for language model alignment. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=H4S4ETc8c9.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=f7OkIurx4b. Survey Certification.

Yoonho Lee, Jonathan Williams, Henrik Marklund, Archit Sharma, Eric Mitchell, Anikait Singh, and Chelsea Finn. Test-time alignment via hypothesis reweighting. *CoRR*, abs/2412.08812, 2024. doi: 10.48550/ARXIV.2412.08812. URL https://doi.org/10.48550/arXiv.2412.08812.

Yafu Li, Xuyang Hu, Xiaoye Qu, Linjie Li, and Yu Cheng. Test-time preference optimization: On-the-fly alignment via iterative textual feedback. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ArifAHrEVD.

Dengcan Liu, Jiahao Li, Zheren Fu, Yi Tu, Jiajun Li, Zhendong Mao, and Yongdong Zhang. Sparserm: A lightweight preference modeling with sparse autoencoder. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(38):32051–32059, Mar. 2026. doi: 10.1609/aaai.v40i38.40476. URL https://ojs.aaai.org/index.php/AAAI/article/view/40476.

Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. LLMs + persona-plug = personalized LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9373–9385, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.461. URL https://aclanthology.org/2025.acl-long.461/.

Feng Luo, Rui Yang, Hao Sun, Chunyuan Deng, Jiarui Yao, Jingyan Shen, Huan Zhang, and Hanjie Chen. Rethinking diverse human preference learning through principal component analysis. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19857–19870, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1019. URL https://aclanthology.org/2025.findings-acl.1019/.

Rajiv Movva, Smitha Milli, Sewon Min, and Emma Pierson. What's in my human feedback? learning interpretable descriptions of preference data. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=sC6A1bFDUt.

Matthew Lyle Olson, Neale Ratzlaff, Musashi Hinck, Man Luo, Sungduk Yu, Chendi Xue, and Vasudev Lal. Probing semantic routing in large mixture-of-expert models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 18263–18278, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.991. URL https://aclanthology.org/2025.findings-emnlp.991/.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

Silviu Pitis, Ziang Xiao, Nicolas Le Roux, and Alessandro Sordoni. Improving context-aware preference modeling for language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 70793–70827. Curran Associates, Inc., 2024. doi: 10.52202/079017-2262. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/82acbbc04435f6c1e7f656b1cbe4ad82-Paper-Conference.pdf.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 52516–52544. Curran Associates, Inc., 2024. doi: 10.52202/079017-1664. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/5e1c255653eb98cef13f45b2d337c882-Paper-Conference.pdf.

Shanghaoran Quan. DMoERM: Recipes of mixture-of-experts for effective reward modeling. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7006–7028, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.418. URL https://aclanthology.org/2024.findings-acl.418/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

Jingyan Shen, Jiarui Yao, Rui Yang, Yifan Sun, Feng Luo, Rui Pan, Tong Zhang, and Han Zhao. MiCRo: Mixture modeling and context-aware routing for personalized preference learning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17447–17463, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.882. URL https://aclanthology.org/2025.emnlp-main.882/.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1690–1712, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.89. URL https://aclanthology.org/2025.findings-emnlp.89/.

Hanshi Sun, Momin Haider, Ruiqi Zhang, Huitao Yang, Jiahao Qiu, Ming Yin, Mengdi Wang, Peter Bartlett, and Andrea Zanette. Fast best-of-n decoding via speculative rejection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=348hfcprUs.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. Arithmetic control of LLMs for diverse user preferences: Directional preference alignment with multi-objective rewards. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8642–8655, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.468. URL https://aclanthology.org/2024.acl-long.468/.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10582–10592, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.620. URL https://aclanthology.org/2024.findings-emnlp.620/.

Xiaobao Wu. A comprehensive survey on learning from rewards for large language models: Reward models and learning strategies. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 17847–17875, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.970. URL https://aclanthology.org/2025.findings-emnlp.970/.

Xinle Wu and Yao Lu. Reward model routing in alignment. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=i3OKIHSsHC.

Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. Regularizing hidden states enables learning generalizable reward model for llms. In *Advances in Neural Information Processing Systems*, 2024.

Xingyi Yang, Constantin Venhoff, Ashkan Khakzar, Christian Schroeder de Witt, Puneet K. Dokania, Adel Bibi, and Philip Torr. Mixture of experts made intrinsically interpretable. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=6QERrXMLP2.

Michael JQ Zhang, Zhilin Wang, Jena D. Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. Diverging preferences: When do annotators disagree and do models know? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=qWgAAVhoXb.

Shuyi Zhang, Wei Shi, Sihang Li, Jiayi Liao, Tao Liang, Hengxing Cai, and Xiang Wang. Interpretable reward model via sparse autoencoder. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(41):34808–34816, Mar. 2026. doi: 10.1609/aaai.v40i41.40783. URL https://ojs.aaai.org/index.php/AAAI/article/view/40783.

Siyan Zhao, John Dang, and Aditya Grover. Group preference optimization: Few-shot alignment of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DpFeMH4l8Q.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10586–10613, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.630. URL https://aclanthology.org/2024.findings-acl.630/.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593, 2019. URL http://arxiv.org/abs/1909.08593.

Thomas P Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. PersonalLLM: Tailoring LLMs to individual preferences. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2R7498e2Tx.

A Details on Experimental Setup

SHP The SHP dataset is curated from Reddit posts and contains 18 subsets featuring different domains. Following the recommended practice from the dataset curator, we use only examples where both chosen and rejected responses receive more than 3 likes. For training and evaluation, we exclude the *explainlikeimfive* and *changemyview* subsets, as their examples contain special prefixes that can easily leak the domain information. For examples replying to the same input, we only sample one pair of responses for evaluation. In the end, we use 36000/4000/1497 examples for train/val/test sets, respectively. The domains, their groupings, and numbers in the test set can be found in Table 6.

Table 6: **Domain composition and number of test examples for each category in SHP.**

Category	Domain	# Examples in the Test Set
Academic	askphysics, askscience, askengineers, askhistorians, askphilosophy, asksocialscience, askanthropology	617
Professional	askdocs, askvet, askhr, legaladvice, askacademia, askcarguys	384
Cooking	askbaking, askculinary	256
Fiction	asksciencefiction	240

RPR Since each example in RPR is paired with two attributes, under which the preference is swapped, we convert each example to two attribute-conditioned examples with context during training. Specifically, following Shen et al. (2025), we append the criteria corresponding to target attributes to the inputs and assign labels accordingly. We use the complete training set with all 40 attributes for reward model training, resulting in 18300/2034 train/validation examples. Some examples of attributes and their corresponding context-dependent criteria are provided below:

- Linguistic Creativity: The response should demonstrate linguistic creativity by incorporating poetry or quotations attributed to the Sufi figure, enhancing the cultural and emotional understanding of their teachings.
- Detail and Elaboration: Provides a detailed step-by-step guide on how to build the app.
- Clarity and Conciseness: Provides a precise and succinct explanation of the bug, focusing on its technical aspects and why it violates design guidelines.

For evaluation robustness, we conduct attribute steering on the attributes that appear over 50 times in the test set, and personalization on the seven of these with the most examples. The attributes appearing over 50 times are: *Linguistic Creativity, User Friendliness, Humor and Entertainment Value, Scientific Rigor, Narrative and Storytelling Quality, Creativity and Originality, Factual Accuracy, Innovativeness, Pedagogical Effectiveness, Economic Feasibility, Technical Complexity, Interdisciplinary Approach, Empathy and Emotional Intelligence, Contextual Relevance, Clarity and Conciseness, User Experience, Practical Application*.

We perform paired t-tests on flip rates across all tested attributes to assess the statistical significance of the improvements.

700K We randomly sample 36000/4000/5000 examples to create the train/validation/test sets for 700K.

License We use publicly available datasets and models for research purposes and cite the original sources. SHP is derived from Reddit data under the Reddit API Terms of Use and does not provide a standard open-source license. RPR is released under CDLA-Permissive-2.0. For preference-700K, we could not identify an explicit license on the dataset card; therefore, we report the license of the official accompanying RLHF Workflow repository, which is released under Apache-2.0. For PersonalLLM, although the Hugging Face metadata lists CC BY 4.0, the paper states that the dataset is released under CC BY-NC 4.0; we therefore conservatively treat it as non-commercial research data. The GRM-Llama3.2-3B reward model and Qwen3.6 models used for expert interpretation and evaluation are released under Apache-2.0. We do not redistribute the original datasets or model weights.

Training Details The router is implemented as a one-hidden-layer MLP with 128 hidden units, and each expert is modeled as a single linear head on top of the backbone model representations. For a fair comparison, we use the same training configuration for all models and baselines. We set the learning rate to 0.002/0.0005 for experts and router parameters, respectively, with a default AdamW optimizer and 5% warmup steps. A dropout rate of 0.2 and jitter noise of 0.25 are applied to the router in MoE models to avoid routing collapse.

We train the models for 3 epochs with a batch size of 16 on one H100 GPU, as we find sufficiently long training is necessary for MoE to develop reasonable specialization.

Evaluation Metrics We provide mathematical descriptions of the evaluation metrics we used in the experiments.

• **Category Coverage** Let $\mathcal{C}$ denote the set of ground truth categories. For each expert k, we assign a category label $\hat{c}_k$ based on the majority category among its top-activating validation examples. Category coverage is defined as:

$$\text{Coverage} = \frac{|\{c \in \mathcal{C} : \exists k, \hat{c}_k = c\}|}{|\mathcal{C}|}. \tag{7}$$

• **Expert Purity** Let $\ell(x) \in \mathcal{C}$ denote the ground truth category of input x. For each expert k, let $\mathcal{T}_k^{(n)}$ be the set of n test examples with the largest routing weights $\pi_{\phi,k}(x)$. Given the category label $\hat{c}_k$ assigned to expert k, expert purity is defined as:

$$\text{Purity}_k = \frac{1}{|\mathcal{T}_k^{(n)}|} \sum_{x \in \mathcal{T}_k^{(n)}} \mathbf{1}\{\ell(x) = \hat{c}_k\}. \tag{8}$$

The overall expert purity score is averaged across experts:

$$\text{Purity} = \frac{1}{K} \sum_{k=1}^{K} \text{Purity}_k. \tag{9}$$

• **Expert Specialization** Let $\mathcal{D}_c$ denote the set of test examples belonging to category c, and let $\text{Acc}_j(c)$ denote the accuracy of expert j on $\mathcal{D}_c$. For each expert k with assigned category $\hat{c}_k$, we define its specialization score as the difference between its accuracy on $\hat{c}_k$ and the average accuracy of all experts on the same category:

$$\text{Spec}_k = \text{Acc}_k(\hat{c}_k) - \frac{1}{K} \sum_{j=1}^{K} \text{Acc}_j(\hat{c}_k). \tag{10}$$

In settings without ground truth categories, we instead evaluate specialization on the examples most strongly associated with each expert. For each expert k, let $\mathcal{T}_k^{(m\%)}$ denote the top m% test examples with the largest routing weights $\pi_{\phi,k}(x)$. Let $\text{Acc}_j(\mathcal{S})$ denote the accuracy of expert j on a set of examples $\mathcal{S}$. We define the specialization score of expert k as:

$$\text{Spec}_k = \text{Acc}_k(\mathcal{T}_k^{(m\%)}) - \frac{1}{K} \sum_{j=1}^{K} \text{Acc}_j(\mathcal{T}_k^{(m\%)}). \tag{11}$$

The overall expert specialization score is averaged across experts:

$$\text{Spec} = \frac{1}{K} \sum_{k=1}^{K} \text{Spec}_k. \tag{12}$$

• **Desired Flip Rate** Let $\mathcal{A}$ denote the set of evaluated attributes and $\mathcal{W}$ denote the set of target expert weights. For each attribute $a \in \mathcal{A}$, let $\mathcal{D}_a$ be its test set and let k_a be the target expert selected for this attribute. Given a target weight $w \in \mathcal{W}$, we define the intervened routing distribution $\tilde{\pi}_\phi^{(k_a,w)}(x)$ by setting the weight of expert k_a to w and rescaling the remaining expert weights to sum to $1 - w$.

Let $\hat{y}^{(k_a,w)}(x)$ denote the response selected by the model under this intervention, and let $y_a^+(x)$ denote the response preferred under attribute a. We first identify the examples that are misclassified when the target expert is removed:

$$\mathcal{M}_a = \left\{ x \in \mathcal{D}_a : \hat{y}^{(k_a,0)}(x) \neq y_a^+(x) \right\}. \tag{13}$$

The desired flip rate for attribute a under target expert weight w is:

$$\text{DFR}(a, w) = \frac{\sum_{x \in \mathcal{M}_a} \mathbf{1}\left\{ \hat{y}^{(k_a,w)}(x) = y_a^+(x) \right\}}{|\mathcal{M}_a|}. \tag{14}$$

The overall desired flip rate is averaged over all evaluated attributes and target expert weights:

$$\text{DFR} = \frac{1}{|\mathcal{A}|\,|\mathcal{W}|} \sum_{a \in \mathcal{A}} \sum_{w \in \mathcal{W}} \text{DFR}(a, w). \tag{15}$$

• **Undesired Flip Rate** For each attribute a, we define the set of initially correct examples under target expert removal as:

$$\mathcal{C}_a = \left\{ x \in \mathcal{D}_a : \hat{y}^{(k_a,0)}(x) = y_a^+(x) \right\}. \tag{16}$$

The undesired flip rate under target weight w is then:

$$\mathrm{UFR}(a, w) = \frac{\sum_{x \in \mathcal{C}_a} \mathbf{1}\left\{\hat{y}^{(k_a, w)}(x) \neq y_a^+(x)\right\}}{|\mathcal{C}_a|}. \tag{17}$$

• **Description Fidelity** For each expert k, let d_k denote its natural language description, and let $\mathcal{S}_k$ be the set of sampled test examples used for evaluation. For each example $x \in \mathcal{S}_k$, we ask an LLM to judge whether x matches d_k, producing a binary relevance judgment $z_k(x) \in \{0, 1\}$. We then compute the Pearson correlation between these judgments and the corresponding router activations $\pi_{\phi,k}(x)$:

$$\mathrm{DF}_k = \mathrm{corr}_{x \in \mathcal{S}_k}\left(z_k(x), \pi_{\phi,k}(x)\right). \tag{18}$$

The overall description fidelity score is averaged across experts:

$$\mathrm{DF} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{DF}_k. \tag{19}$$

Expert Interpretation When computing the description fidelity of expert interpretations, since only a small fraction of examples strongly activate any given expert, we stratify the sample by drawing 100 examples from the top 10% of activations and 100 examples from the remaining 90%.

The prompts that we use for interpreting experts and evaluating interpretability are adapted from Movva et al. (2026) and are provided in Figure 4 and Figure 5.

Test-Time Personalization We use the test-time weight adjustment method from Lee et al. (2024) for HyRe and implement the Hedge algorithm from Shen et al. (2025) for fine-tuning routers in MoE models. For Hedge, we use $\tau = 0.001$, following the recommendation of the original paper. All models are trained for 10 epochs using a batch size of 32 and a learning rate of 0.01 for adaptation.

B Effect of the Number of Experts

We study the effect of the number of experts by training sparse MoE models with $K \in \{10, 20, 30, 40, 50\}$ on 700K and evaluating their interpretability. We conduct this analysis on the real-world dataset rather than the controlled settings, because the pre-defined labels in the controlled experiments capture only coarse latent structures and may not fully reflect the finer-grained patterns discovered by the model.

Figure 6 shows how description fidelity and expert specialization change with the number of experts (single run with seed=42). Description fidelity increases from $K = 10$ to $K = 30$, where it reaches its peak, and then remains relatively stable. This suggests that increasing the number of experts initially helps the router discover more semantically coherent input clusters, but adding experts beyond this point provides limited additional benefit. We interpret this peak as reflecting the effective granularity of input patterns that can be reliably supported by the data, training objective, and interpretation procedure.

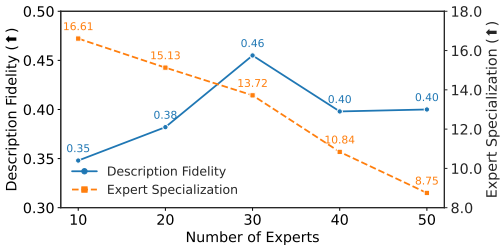

Figure 6: **Effect of the number of experts on interpretability.** Description fidelity peaks at a certain number of experts, while expert specialization consistently decreases as the number of experts increases.

By contrast, expert specialization decreases steadily as K increases. This suggests that larger expert counts may fragment the preference learning signal and weaken the behavioral specialization of individual expert heads. Together, these trends indicate a trade-off between semantic interpretability and functional specialization: router assignments can remain semantically interpretable at larger K, while the corresponding expert heads become less behaviorally specialized. We therefore recommend selecting K based on the complexity of the target data distribution.

Prompt: Describing experts in natural language given top-activating examples

Task Overview
You are a machine learning researcher analyzing a sparse mixture-of-experts reward model.
The model contains a router that assigns each input prompt to one or more experts. You are trying to understand what feature of the input prompt causes a specific expert to receive high routing weight.
Each dataset example consists only of a user prompt. These prompts were selected because this expert was strongly activated for them, meaning the router assigned this expert a high weight.
Your goal is to identify the common feature shared by these prompts, and to describe this feature as a concise, natural language concept.

Instructions
First, inspect each prompt and consider what concrete property it has.
Then, synthesize the single feature that is most consistently shared across the prompts.
The feature should describe the input prompt itself, not the model, the router, or the expert.
For example, the common feature might be:

- "asks for emotionally supportive advice"
- "requests factual explanation of a technical concept"
- "asks for creative writing or brainstorming"
- "contains a safety-sensitive or potentially harmful request"
- "asks for evaluation of competing options"

If the prompts do not share a clear topic, focus on the type of user intent, task format, reasoning style, domain, constraint, or interaction pattern that best explains why the expert fires.

Additional Rules
Rules about the feature you identify:

- The feature should be objective and concrete.
- Err on the side of being concise.
- The feature should be an attribute that could be present in an individual prompt.
- Do not refer to "the expert", "the router", "activation", or "these examples" in the final answer.
- Do not use vague descriptions such as "is helpful", "is complex", or "is high quality".
- Avoid overly broad concepts such as "general question" unless no more specific feature is apparent.
- If the shared feature is a negation, describe it directly, e.g., "does not specify a concrete task or goal".

Prompts

{examples}

Output
Do not output anything besides the feature.
Output exactly one description, starting with "-" and surrounded by quotes.

Your response is:-"

Figure 4: **Prompt used to summarize the common feature of top-activating prompts and describe it in natural language.**

C Ablation Study of Interpretability Regularization

To better understand the effect of each interpretability regularization term on model behavior, we conduct a comprehensive ablation study using a grid search over $\lambda_{ls} \in [0.0, 0.5, 1.0]$, $\lambda_{gb} \in [0.0, 0.5, 1.0]$, and $\lambda_{div} \in [0.0, 1.0, 2.0]$ on SHP with $K = 5$ experts. The resulting task performance, routing pattern, expert behavior, and interpretability metrics are reported in Table 7 (single run with seed=42).

For ease of visualization, we highlight in **bold** the models that do not suffer from collapse. These models exhibit low local entropy and high global entropy, indicating that the MoE does not rely on a single expert nor distribute inputs uniformly across experts. We observe that the interaction between λ_{ls} and λ_{gb} largely determines whether the trained MoE model collapses. These two hyperparameters influence routing patterns in the intended manner: when the model uniformly utilizes all experts (i.e., high local and global entropy), increasing λ_{ls} encourages more confident routing de-

Prompt: Judging whether an example matches expert description

```
You are a binary classifier.

Task:
Determine whether the input text expresses the feature described below.

Feature description:
{description}

Input text:
{input_text}

Decision rule:
Answer YES if the input text clearly matches or expresses the feature description.
Answer NO otherwise.

Output format:
Return exactly one token: YES or NO.

Answer:
```

Figure 5: **Prompt used to judge whether an input example matches a given natural language feature description.**

Table 7: **Ablation study of interpretability regularization terms on SHP.** Models that do not suffer from collapse are marked in **bold**. λ_{ls} and λ_{gb} mainly determine whether the MoE model collapses, while λ_{div} affects model interpretability and expert specialization.

λ_{ls}	λ_{gb}	λ_{div}	Accuracy (↑)	Routing Entropy (↓)	Global Entropy (↑)	Expert Correlation (→ 0)	Expert Purity (↑)	Expert Specialization (↑)
0.00	0.00	0.00	70.67	0.991	0.993	0.556	74.40	-1.00
0.00	0.00	1.00	70.70	0.133	0.139	0.024	80.00	1.12
0.00	0.00	2.00	71.10	0.014	0.015	0.003	70.00	0.47
0.00	0.50	0.00	71.17	1.000	1.000	0.493	43.20	-0.27
0.00	0.50	1.00	71.54	1.000	1.000	0.141	46.40	0.62
0.00	0.50	2.00	71.49	1.000	1.000	0.126	49.20	-0.07
0.00	1.00	0.00	71.24	1.000	1.000	0.520	56.80	-0.20
0.00	1.00	1.00	70.79	1.000	1.000	0.153	47.60	-0.91
0.00	1.00	2.00	71.92	1.000	1.000	0.110	47.20	0.84
0.50	0.00	0.00	70.24	0.000	0.000	0.743	54.80	-0.57
0.50	0.00	1.00	71.44	0.000	0.000	-0.028	46.80	-0.71
0.50	0.00	2.00	71.19	0.000	0.000	-0.027	44.00	-0.93
0.50	**0.50**	**0.00**	**71.39**	**0.087**	**0.984**	**0.782**	**43.20**	**0.06**
0.50	**0.50**	**1.00**	**68.55**	**0.131**	**0.983**	**0.016**	**75.20**	**4.66**
0.50	**0.50**	**2.00**	**69.62**	**0.200**	**0.986**	**0.011**	**85.60**	**5.77**
0.50	**1.00**	**0.00**	**71.40**	**0.084**	**0.989**	**0.746**	**46.80**	**0.34**
0.50	**1.00**	**1.00**	**66.96**	**0.293**	**0.960**	**0.067**	**54.00**	**4.12**
0.50	**1.00**	**2.00**	**69.45**	**0.333**	**0.917**	**0.033**	**67.20**	**3.34**
1.00	0.00	0.00	70.52	0.000	0.000	0.673	46.40	-0.69
1.00	0.00	1.00	70.64	0.000	0.000	-0.026	54.00	-1.23
1.00	0.00	2.00	70.76	0.000	0.000	0.006	34.00	-0.76
1.00	**0.50**	**0.00**	**70.31**	**0.214**	**0.909**	**0.783**	**49.60**	**-0.06**
1.00	0.50	1.00	70.05	0.142	0.148	0.044	80.40	1.10
1.00	0.50	2.00	70.54	0.161	0.166	0.007	92.00	1.35
1.00	**1.00**	**0.00**	**71.17**	**0.083**	**0.983**	**0.746**	**47.20**	**0.38**
1.00	**1.00**	**1.00**	**69.31**	**0.088**	**0.987**	**0.072**	**47.20**	**3.71**
1.00	**1.00**	**2.00**	**66.15**	**0.104**	**0.997**	**0.009**	**54.80**	**3.98**

cisions; conversely, when the model collapses to a single expert (i.e., low local and global entropy), increasing λ_{gb} encourages the model to utilize a broader set of experts.

While λ_{ls} and λ_{gb} primarily control router behavior, λ_{div} directly affects model interpretability and expert specialization. Increasing λ_{div} encourages more diverse expert outputs, leading to stronger expert specialization. At the same time, it imposes additional pressure on the router to distribute examples in a more semantically meaningful manner, resulting in higher expert purity.

Overall, although the configuration we study involves three regularization terms and their corresponding hyperparameters, the ablation study shows that they serve distinct roles, allowing users to tune them based on observed model behavior. Experiments in Appendix D further demonstrate the robustness of our hyperparameter choices across different backbone models and datasets. When training sparse MoE reward models on new datasets, we therefore recommend starting from the hyperparameter configuration that works well on known data, then tuning λ_{ls} and λ_{gb} to ensure stable training. λ_{div} can subsequently be adjusted according to the desired degree of interpretability.

Table 8: **Interpretability results across different backbone models and datasets.** Best scores are marked in **bold**.

(a) 700K with Qwen3-8B as backbone.

Method	Accuracy (↑)	Description Fidelity (↑)	Expert Specialization (↑)
Single Reward	82.02	-	-
Vanilla MoE	82.56	0.23	1.55
MiCRo	**83.02**	0.01	-0.09
Random MoE	-	-0.00	-0.33
Sparse MoE	82.64	**0.41**	**16.75**

(b) LMArena with Llama3.2-3B as backbone.

Method	Accuracy (↑)	Description Fidelity (↑)	Expert Specialization (↑)
Single Reward	**64.56**	-	-
Vanilla MoE	61.10	0.08	0.73
MiCRo	60.18	0.07	0.91
Random MoE	-	-0.01	0.63
Sparse MoE	60.22	**0.23**	**7.88**

Table 9: **RPR personalization results with Qwen3-8B-based reward models trained on 700K.** Best scores are marked in **bold**.

Method	Linguistic Creativity	User Friendliness	Humor & Entertainment	Scientific Rigor	Narrative & Storytelling	Creativity & Originality	Factual Accuracy	Overall
Single Reward	20.19	31.46	27.38	83.33	33.75	26.39	83.10	43.66
HyRe	25.64 (+7.37)	30.71 (+0.37)	30.16 (+6.35)	**90.48 (+1.19)**	37.92 (+4.17)	35.65 (+6.48)	84.51 (+0.00)	47.87 (+3.70)
Vanilla MoE	53.20 (+33.01)	49.06 (+15.35)	56.75 (+31.75)	82.14 (-3.57)	57.92 (+26.67)	72.68 (+46.29)	**89.67** (+0.94)	65.92 (+21.49)
MiCRo	22.44 (+3.21)	39.70 (+8.24)	25.40 (+1.59)	88.10 (-2.38)	30.83 (+2.08)	34.72 (+8.33)	85.45 (-0.47)	46.66 (+2.94)
Sparse MoE	**83.97 (+56.09)**	**61.05 (+29.59)**	**97.22 (+72.22)**	85.71 **(+1.19)**	**67.50 (+30.00)**	**81.02 (+57.41)**	84.98 **(+6.11)**	**80.21 (+36.09)**

D Generalization Results

We train MoE models under two additional settings: using a different pretrained reward model with a distinct architecture, Qwen3-8B (Skywork-Reward-V2-Qwen3-8B), on the 700K dataset, and using the same Llama-3.2-3B model on a different preference dataset, LMArena. We intentionally use the same hyperparameters across settings to evaluate the transferability of our hyperparameter choices.

Table 8 reports the interpretability of these models, compared with other methods implemented using the same respective base models and datasets. Consistent with the main experiments, sparse MoE exhibits the strongest interpretability in both settings, achieving significantly higher description fidelity and expert specialization. These results support the generalizability of sparse MoE's interpretability across different base models and datasets, as well as the robustness of our hyperparameter selection.

Tables 9 and 10 further demonstrate that sparse MoE maintains strong personalization performance across base models and datasets. With the Qwen backbone, it substantially outperforms the baselines in both post-adaptation performance and personalization gains. When trained on LMArena, sparse MoE still achieves the largest personalization improvements. These results suggest that our core finding—that interpretable and specialized experts facilitate effective personalization—generalizes across settings.

E Comparing MoE and SAE Reward Models

In addition to the quantitative results, we also provide a qualitative comparison of MoE- and SAE-based interpretations. In Table 11, we sample five representative features or experts from two SAE-based methods, SARM and SparseRM (Liu et al., 2026), and compare them against sparse MoE interpretations.

We observe that SAE-based models learn qualitatively different types of features from sparse MoE. This difference may be explained by their training paradigms. Both SparseRM and SARM rely on SAE encoders pretrained on large-scale corpora, and then learn reward weights over the extracted features on 700K. Since SparseRM uses a token-level SAE, its features tend to capture lexical, formatting, or local syntactic patterns. SARM operates at the sequence level and therefore captures more response-level patterns, some of which are preference-relevant. However, despite their computationally expensive pretraining, both methods may suffer from a mismatch between SAE pretraining and the downstream preference-modeling objective. As a result, their interpretations are less specific to the target dataset and yield worse quantitative interpretability results.

Table 10: **RPR personalization results with Llama3.2-3B-based reward models trained on LMArena.** Best scores are marked in **bold**.

Method	Linguistic Creativity	User Friendliness	Humor & Entertainment	Scientific Rigor	Narrative & Storytelling	Creativity & Originality	Factual Accuracy	Overall
Single Reward	83.65	49.44	89.29	28.57	85.00	79.17	25.35	62.92
HyRe	86.10 (+12.06)	71.16 **(+14.98)**	90.25 (+0.96)	**69.22 (+28.74)**	80.33 (+5.33)	82.65 (+15.98)	80.10 (+40.66)	**79.97** (+16.96)
Vanilla MoE	88.46 (+14.42)	60.67 (+2.24)	**93.26 (+7.55)**	61.11 (+23.01)	86.25 (+13.75)	86.57 (+21.29)	74.65 (+32.40)	78.71 (+16.38)
MiCRo	87.82 (+18.59)	65.54 (+5.99)	**93.26** (+6.36)	49.60 (+11.50)	**87.50** (+16.25)	**87.50 (+23.61)**	63.38 (+23.94)	76.37 (+15.18)
Sparse MoE	**89.42 (+25.00)**	**72.28** (+9.36)	88.89 (+6.75)	63.89 (+26.99)	77.08 **(+20.83)**	81.02 (+22.69)	**81.69 (+40.84)**	79.18 **(+21.78)**

Table 11: **Representative feature/expert interpretations in sparse MoE and SAE-based reward models.** These examples reveal substantial qualitative differences in the types of patterns captured by the two model families.

Sparse MoE	**SARM**	**SparseRM**
asks for advice, feedback, or discussion on a personal, professional, or academic challenge	provides a brief, direct, or conversational response to a user prompt	textual structures, particularly punctuation and formatting indicators
asks to solve a procedural task involving numerical or list manipulation	provides a structured, step-by-step explanation or guide	specific verbs or actions associated with legal or formal contexts
asks for a summary or extraction of information from a provided text	provides a direct, definitive answer to a factual or technical question	patterns and structures in data or questions
asks for a direct answer to a multiple-choice or short-answer question	provides an incorrect, irrelevant, or hallucinated response to the prompt	specific names, titles, or identifiers related to individuals and organizations
requests assistance with illegal, harmful, or sexually explicit activities	provides a step-by-step explanation or structured breakdown of the answer	elements related to programming or coding structures

F Full Results

Category Recovery Results Table 12 and Table 13 present the overall performance of our sparse MoE model on SHP (single run with seed 42), together with the accuracy and routing weights of each expert. We observe that the global entropy within each category is lower than that on the full dataset, indicating semantically reasonable routing patterns at the category level. Examining the expert distributions further, we find that most categories assign more than 50% of their routing weight to a single expert. Moreover, these dominant experts typically achieve strong performance within their corresponding categories. Together, these results suggest that the sparse MoE reward model learns meaningful routing patterns and behaviorally specialized experts.

We further analyze the performance drop of the sparse MoE reward model compared to the single-head baseline. The largest degradation occurs in the academic and professional categories, where the best-performing experts do not receive the highest routing weights. We therefore attribute the accuracy drop of sparse MoE mainly to a mismatch between routing weights and expert abilities. This mismatch likely arises because these two categories are broad and heterogeneous, making it more difficult for the router to assign examples to the most suitable experts.

Expert Descriptions Table 14 presents the natural language descriptions of all experts in the sparse MoE model trained on 700K examples. Most experts specialize in a concrete type of question or task, exhibiting high description fidelity while also specializing behaviorally in the examples routed to them. Despite nuanced differences between experts, many focus on recurring categories such as feedback-seeking, safety-related queries, mathematics, code generation, and question answering. These expert patterns partially reflect the underlying structure and distribution of the training data.

Table 12: **Category-level evaluation results on SHP.** Scores in parentheses denote performance comparison against the single-head reward model.

Domain	# Examples	Accuracy (↑)	Routing Entropy (↓)	Global Entropy (↑)	Expert Correlation (→ 0)
Academic	617	70.50 (-7.06)	0.252	0.883	0.079
Professional	384	70.31 (-4.43)	0.149	0.733	0.038
Cooking	256	66.41 (-0.39)	0.156	0.568	-0.006
Fiction	240	71.25 (+0.84)	0.198	0.704	0.046
Overall	1497	69.62 (-0.90)	0.200	0.986	0.011

Table 13: **Expert accuracy and average routing weights on SHP.** Each cell reports expert accuracy / routing weight.

Domain	Expert 0	Expert 1	Expert 2	Expert 3	Expert 4
Academic	65.80 / **35.56**	67.59 / 25.93	62.40 / 10.46	55.92 / 3.76	**69.53** / 24.28
Professional	58.85 / 10.98	52.34 / 6.91	67.45 / **55.40**	52.34 / 2.06	**68.75** / 24.66
Cooking	61.33 / 7.18	61.72 / 21.17	49.61 / 1.19	**68.36** / **68.28**	59.77 / 2.18
Fiction	63.33 / 15.81	56.25 / 4.33	66.25 / 28.57	46.25 / 0.08	**68.33** / **51.21**
Overall	62.86 / 21.23	60.86 / 16.78	62.12 / 23.31	55.58 / 13.77	**67.47** / **24.92**

Table 14: **Natural language descriptions of all experts in the sparse MoE model.**

ID	Natural Language Description	Description Fidelity (↑)	Expert Specialization (↑)
0	asks for advice, feedback, or discussion on a personal, professional, or academic challenge	0.43	22.08
1	asks for natural language inference or textual entailment classification	0.26	22.48
2	requests the generation of computer code or technical scripts	0.45	21.53
3	asks for troubleshooting advice regarding a specific problem in cooking, baking, or home maintenance	0.32	15.84
4	asks for personal advice or recommendations	0.50	6.80
5	contains a request for harmful, illegal, or policy-violating content	0.43	7.97
6	asks to solve a procedural task involving numerical or list manipulation	0.51	25.98
7	asks for an opinion, consensus, or general advice on a subjective or non-factual topic	0.41	6.20
8	asks for a summary or extraction of information from a provided text	0.45	19.35
9	asks for general information, advice, or explanation on a broad topic	0.62	8.59
10	asks for a direct answer to a multiple-choice or short-answer question	0.36	20.56
11	asks for technical advice, architectural design, or problem-solving in software engineering and IT	-0.19	19.49
12	requests technical implementation details, system design, or professional strategy plans	0.39	18.69
13	requests assistance with illegal, harmful, or sexually explicit activities	0.62	3.35
14	asks for legal or procedural advice regarding a specific real-world situation	0.33	24.32
15	asks to solve a competitive mathematics problem	0.42	32.00
16	asks for a list of examples, suggestions, or items	0.30	8.60
17	asks a question containing a factual error, category mistake, or logical impossibility	0.30	3.92
18	requests assistance with illegal, harmful, or unethical activities	0.36	6.76
19	asks a question containing a factual error, logical fallacy, or absurd premise	0.35	8.12

We also observe that experts with similar activation patterns, measured by high correlations in their activations across examples, tend to receive similar interpretations. For instance, expert 4 exhibits strong activation correlations with expert 7 (0.95) and expert 16 (0.93), and all three are identified as advice-seeking experts. Similarly, safety-related experts also display highly aligned activation patterns, as illustrated by the 0.96 correlation between experts 5 and 18.

These findings suggest the potential for hierarchical expert interpretation and for reducing expert redundancy through additional regularization on router activation correlations, which we leave for future work.

G Details on Human Study

We recruited five annotators with at least a master's degree in computer science, computational linguistics, or linguistics to evaluate the interpretability of sparse MoE and the baseline models. Annotators were first asked to rate pattern coherence after seeing the top-activating examples for each expert. They were then shown the LLM-generated natural language description and asked to rate its quality.

The ratings show fair to moderate inter-annotator agreement. Sparse MoE achieves statistically significant improvements in pattern coherence over all baselines ($p < 0.05$), while the improvements in description quality are not statistically significant. Statistical significance is computed with a paired t-test.

The annotation instructions and rating scales for pattern coherence and description quality are provided below.

Pattern Coherence Pattern coherence measures whether the input prompts in the same group share a consistent and recognizable pattern. The shared pattern could be a specific task, shared topic, semantic feature, input format, writing style, user intent, domain, or another interpretable property. A high score means that the prompts appear meaningfully related and could plausibly be summarized by a common feature. A low score means that the prompts appear random, unrelated, or too diverse to be summarized by a single coherent pattern.

1. **Not Coherent:** The prompts appear unrelated, random, or inconsistent. No clear shared task, topic, feature, format, style, intent, or domain can be identified.
2. **Weakly Coherent:** A few prompts share some similarities, but the overall group is noisy and difficult to summarize with a single pattern.
3. **Moderately Coherent:** The prompts share a partially recognizable pattern, but there are also several exceptions or mixed themes.
4. **Mostly Coherent:** Most prompts share a clear common task, topic, feature, format, style, intent, domain, or other interpretable property, with only minor exceptions.
5. **Highly Coherent:** The prompts are strongly and consistently related. They clearly reflect a common interpretable concept.

Description Quality Description quality measures how accurately the provided natural language description captures the common pattern shared by the input prompts. This pattern may involve a specific task, shared topic, semantic feature, input format, writing style, user intent, domain, or another interpretable property. A high score means that the description correctly summarizes the main shared pattern among the prompts without adding misleading or unsupported claims. A low score means that the description is inaccurate, too vague, inconsistent with the prompts, or misses the main pattern.

1. **Poor:** The description does not match the prompts, misses the main pattern, or describes a feature that is mostly absent.
2. **Limited:** The description captures only a small part of the pattern, or is mostly vague, incomplete, or partially incorrect.
3. **Adequate:** The description captures some important shared features, but it is still vague, misses other key aspects or includes noticeable inaccuracies.
4. **Good:** The description accurately captures the main pattern of the prompts, with only minor omissions or minor overgeneralizations.
5. **Excellent:** The description clearly and accurately summarizes the shared features of the prompts, with no meaningful inaccuracies.

Human Study Ethics and Compensation Before participating in the study, annotators were informed about how their data would be used and were told that their annotations would be deleted upon request. The study took approximately 20 minutes on average, and annotators were compensated at an hourly rate comparable to the local average hourly salary for researchers. The human study protocol has been approved by the institutional ethics board.

H Effect of Adaptation Set Size

Figure 7 shows the attribute-level personalization performance of the sparse MoE model under different adaptation set sizes. Results are averaged over the seven attributes and three independently sampled adaptation sets. We observe that, even with only 5 adaptation examples, the model improves accuracy by approximately 14 points. As expected, adaptation performance further improves with larger adaptation sets and begins to saturate at around 100 examples.

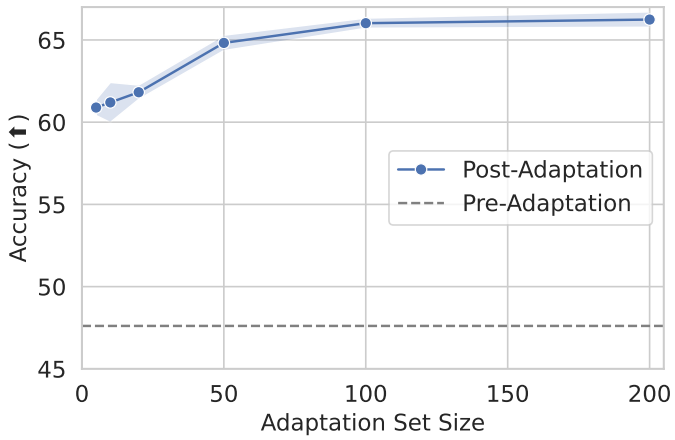

Figure 7: **Personalization results of sparse MoE with different adaptation set sizes on RPR.** Even with only 5 adaptation examples, sparse MoE achieves strong personalization performance. Performance further improves with larger adaptation sets before saturating at around 100 examples.

Table 15: **Experts most strongly upweighted after adaptation for selected attributes.** The experts align semantically with the adapted attributes in an interpretable way.

Attribute	Most Upweighted Expert
Factual Accuracy	asks for a summary or extraction of information from a provided text
Pedagogical Effectiveness	asks for advice, feedback, or discussion on a personal, professional, or academic challenge
Innovativeness	asks a question containing a factual error, logical fallacy, or absurd premise

I Inspecting Preference Adaptation

Leveraging the interpretability and specialization of experts in sparse MoE, we further analyze the model's adaptation to individual preferences by examining changes in expert weights after adaptation. Table 15 reports, for a subset of attributes, the expert whose weight increases the most after adaptation. We observe that the upweighted experts often show plausible semantic associations with the target attributes. Overall, post-adaptation expert weight shifts provide a useful diagnostic signal for inspecting preference adaptation.

J User-Level Personalization

We use PersonalLLM for user-level adaptation. Following prior work (Bose et al., 2025; Choi et al., 2025), we model each user as a weight vector over ten reward models. Adaptation and evaluation examples are then constructed from the highest- and lowest-ranked responses according to user-specific weights. We sample 5 users from a Dirichlet distribution with $\alpha = 0.3$, and use $n = 50$ examples for adaptation. Evaluation is conducted on 200 synthetic examples. Results are averaged over three independently sampled adaptation sets.

Table 16 shows that our sparse MoE reward model also achieves the strongest performance in user-level adaptation. Consistent with the attribute-level adaptation results in Table 5, sparse MoE attains both the best post-adaptation performance and the largest improvements over the non-adapted model. By contrast, the two MoE baselines, vanilla MoE and MiCRo, fail to adapt effectively to new users, with performance degrading after adaptation. The static-weight ensemble model HyRe performs better than the MoE baselines, likely because its simpler adaptation mechanism is more robust in low-data settings.

Table 16: **Personalization results on PersonalLLM.** Scores in parentheses denote the change after adaptation. Higher scores are better. Sparse MoE also demonstrates strong performance in user-level adaptation.

Method	User1	User2	User3	User4	User5	Overall
Single Reward	93.83	94.83	91.50	94.67	87.17	92.40
HyRe	95.50 (+0.83)	95.17 (-0.33)	**93.00** (-0.17)	95.33 (-0.50)	**89.50** (+1.00)	93.70 (+0.17)
Vanilla MoE	94.83 (+0.00)	94.67 (-0.83)	90.67 (-2.00)	94.67 (-1.33)	88.83 (+0.50)	92.73 (-0.74)
MiCRo	94.50 (-0.33)	94.67 (-0.83)	90.67 (-2.00)	94.83 (-1.17)	88.17 (-0.16)	92.57 (-0.90)
ArmoRM	68.67 (+0.84)	72.67 (+1.84)	71.50 (**+1.67**)	72.17 (0.00)	67.17 (**+1.85**)	70.44 (+1.24)
Sparse MoE	**95.83 (+4.50)**	**96.17 (+3.50)**	92.67 (+1.34)	**97.33 (+3.33)**	88.50 (-0.33)	**94.10 (+2.47)**

K Personalized Alignment with Adapted Preferences

K.1 Sparse MoE for Downstream Post-Training

To evaluate whether the adapted preferences can support personalized alignment of LLMs, we post-train policies using adapted sparse MoE reward models corresponding to different attributes from RPR. Specifically, we select two attributes, creativity and scientific rigor, which exhibit improved performance after personalization.

For simplicity, we perform GRPO (Shao et al., 2024) post-training on prompts from the RPR training set for 1000 optimization steps. We use an effective batch size of 8 and sample 4 rollouts per prompt. Policies are initialized from Llama3.2-3B-Instruct.

During evaluation, we generate responses for 200 examples from the RPR test set and compare them against a baseline policy trained using the non-adapted reward model. An LLM judge (Qwen3.6-27B) evaluates which response better reflects the target attribute while maintaining higher overall response quality.

Table 17: **Personalized alignment results compared against alignment using a universal reward model.** After personalization, the sparse MoE reward model can effectively guide policies toward the target attributes without negatively affecting overall generation quality.

Alignment Type	Creativity		Scientific Rigor	
	Win Rate (Target Attribute, ↑)	Win Rate (Quality, ↑)	Win Rate (Target Attribute, ↑)	Win Rate (Quality, ↑)
Personalized	**53.0**	**52.5**	**52.0**	**52.5**
Universal	45.5	45.5	44.5	45.0

Results in Table 17 show that post-training policies with personalized reward models better align the policies with the target attributes compared to using a universal preference model. This demonstrates that the personalization improvements of our sparse MoE reward models also generalize to the post-training phase.

K.2 Personalized Reward Models for Pluralistic Alignment

In this work, we mainly focus on one aspect of pluralistic alignment: how to *represent* diverse human values in an interpretable and adaptable reward model. Optimizing policies toward these individualized value representations is a complementary research direction, and here we discuss how an individualized sparse MoE model can support pluralistic alignment in several ways.

First, one can train or fine-tune a separate policy for each user. In Appendix K, we show that an adapted sparse MoE model, trained with only a small number of examples (around 50), can be used as the optimization signal for GRPO to steer a policy toward the target preference direction. This approach is conceptually direct, but computationally expensive.

A more efficient option is to update and maintain only lightweight user-specific components, such as LoRA modules or user embeddings. Related test-time optimization methods (Li et al., 2025) may also reduce the need for full offline policy training.

The adapted reward model can also be used without explicit policy training, for example by guiding a shared universal policy at inference time. Reward-guided decoding methods such as PAD (Chen et al., 2025b), DeAL (Huang et al., 2025), and RAD (Deng & Raffel, 2023) show that reward models or customized reward functions can be used for decoding-time alignment without maintaining separate policies. In our setting, the individualized sparse MoE reward model could provide the user-specific reward signal.

Finally, generate-then-select methods, such as Best-of-N sampling (Ichihara et al., 2025; Sun et al., 2024), offer another simple training-free alternative: a universal policy generates multiple candidate responses, and the individualized reward model selects the response that best aligns with the target user's preferences.

L LIMITATIONS

Our study has certain limitations that should be acknowledged.

First, sparse MoE training introduces additional hyperparameters and tuning complexity. We ablate the number of experts and the regularization terms (Appendices B and C) and show that one configuration remains effective across backbones and datasets (Appendix D). Our results therefore establish that interpretable preference components can be learned under suitable conditions; broader robustness across training configurations remains open.

Second, because the router conditions only on the prompt and not on the responses, router-based interpretations reveal the prompt-identifiable domains in which experts specialize rather than how they evaluate individual responses. Such task- and domain-level specialization is still meaningful for personalization, where users value performance differently across contexts, as our adaptation experiments show.

Third, sparse MoE slightly underperforms other MoE variants on universal preference accuracy. We stress that the two accuracies at stake are different: what sparse MoE gives up is 1.04 points relative to vanilla MoE on the population-average preference function (700K), while what it gains is an 18.63-point advantage over the same architecture on unseen preferences after adaptation (RPR). The trade-off is real but asymmetric: sparse MoE is most appropriate when interpretability and personalization are the objectives. Appendix F analyzes the residual gap, which partly arises from misalignment between router assignments and expert strengths.

Finally, our study covers existing MoE reward models (Wang et al., 2024b; Shen et al., 2025) trained with standard interpretability regularizers; how far expert structure could be pushed by purpose-built architectures or objectives remains open. Our contribution is diagnostic—what these models learn, why conventional training leaves it entangled, and what that entanglement costs—together with the evaluation protocol that supports such claims. We hope both facilitate further work on interpretable preference modeling.