Pluralistic On-Policy Self-Distillation

Yiran Shen [1] **Yu Xia** [1] **Liuyi Yao** [2] **Prithviraj Ammanabrolu** [1]

Abstract

Language feedback often contains multiple valid persona-dependent directions for improvement: a critique may ask a response to match the style of a professional advisor, a travel guide, or an artistic critic. This creates a challenge for pluralistic alignment, where distinct persona-specific feedback signals should be preserved rather than collapsed into a single reward or generic target. We propose Multi-Action-Head On-Policy Self-Distillation (MAH-OPSD), which combines persona-specific feedback with dense token-level on-policy distillation. For each prompt, MAH-OPSD first generates persona-specific rubrics to elicit more targeted critiques than generic feedback criteria. It then trains multiple persona action heads on a shared backbone: each head generates a response from the same prompt, receives its own rubric-guided critique, and distills from a critique-conditioned base model as its teacher. A lightweight router mixes the learned action heads based on the prompt, enabling adaptive response generation at inference time. We validate MAH-OPSD in two persona-faceted settings: a five-persona alignment task with rubric-guided critiques, and a multi-turn tutoring task with two teacher personas, where the per-turn feedback is the student's own reaction rather than an external critique. In both, the action heads specialize by persona and a learned router exposes them as a single adaptive policy, preserving distinct feedback pathways rather than merging all feedback into one generic policy.

1. Introduction

Feedback on large language model (LLM) outputs is often persona-dependent. The same response may be too generic for a professional advisor, too dry for a travel guide, or too narrow for an artistic critic. These judgments are not merely noisy estimates of one hidden reward; they reflect different valid directions for improvement. A central challenge in learning from feedback is therefore not only how to extract supervision from critiques, but also how to preserve the distinctions among them. This challenge is closely related to pluralistic alignment, where models should respect diversity in users, contexts, and response norms rather than optimize only average preference satisfaction (Sorensen et al., 2024; Feng et al., 2024; Castricato et al., 2025; Xie et al., 2025). In this view, disagreement should often be treated as signal rather than noise. This raises a practical training question: how can a single language model used at inference time learn from multiple persona-specific feedback signals without collapsing them into one generic behavior?

Natural-language feedback is a promising signal for persona-dependent alignment because critiques can explain why a response fails, identify missing constraints, and describe how the answer should change. Existing methods use such feedback to refine model outputs or distill feedback-informed behavior into a feedback-free policy (Akyürek et al., 2023; Chen et al., 2023; 2024; Wang et al., 2026; Song et al., 2026). On-policy self-distillation (OPSD) is especially suitable here: the student samples responses under the inference prompt, while a privileged teacher conditioned on feedback provides dense token-level guidance on the states the student actually visits (Agarwal et al., 2024; Xu et al., 2025; Lu & Lab, 2025; Yang et al., 2025). This allows the inference-time model to benefit from natural-language feedback without requiring feedback at test time. However, existing feedback-based distillation methods typically merge all transferred signals into a single policy (Hübotter et al., 2026; Song et al., 2026; Xiao et al., 2026; Yang et al., 2026; DeepSeek-AI, 2026). For persona-dependent feedback, this merging can erase the differences that the model should preserve.

In this paper, we instantiate this problem in a controlled persona-faceted setting, where each persona defines a different response norm and critique style. This setting lets us study whether a model can learn from multiple natural-language feedback signals while keeping them distinct, rather than treating all critiques as supervision for the same generic behavior. We propose Multi-Action-Head On-Policy Self-Distillation (MAH-OPSD), which combines persona-specific feedback with dense token-level on-policy distil-

[1]University of California, San Diego [2]Alibaba Group. Correspondence to: Yiran Shen <jes038@ucsd.edu>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

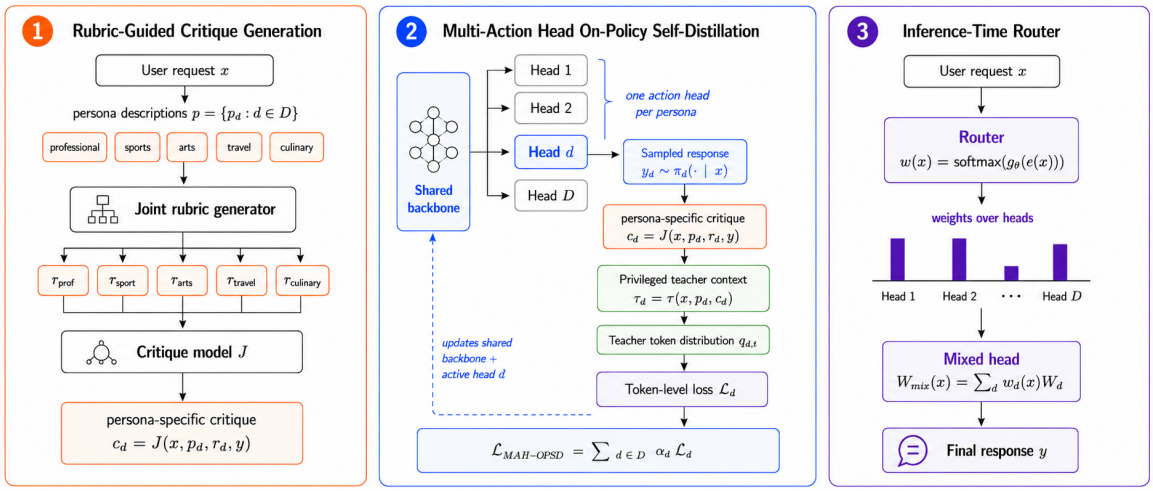

Figure 1. Overview of Multi-Action Head On-Policy Self-Distillation.

lation. The model uses a shared LLM backbone with one action head per persona, following the multi-action-head architecture proposed by (Shen et al., 2025). For each prompt and persona, MAH-OPSD first generates a prompt-specific rubric to guide critique generation, so feedback is tied to the current prompt and persona rather than to generic criteria. The corresponding head then generates an on-policy response under the inference prompt, receives its rubric-guided critique, and distills from a critique-conditioned base model as its teacher. This preserves the dense supervision and on-policy relevance of self-distillation while preventing different persona feedback signals from being forced through the same output head. At inference time, explicit persona labels or critiques may not be available. We therefore train a lightweight router that maps the prompt to a mixture over the learned action heads, enabling adaptive response generation from the prompt alone. We further show that the same recipe applies when the privileged feedback is a naturally occurring interaction signal rather than an LLM critique. In a multi-turn tutoring task whose two personas are contrasting teaching styles (a direct-instruction persona and an interactive one), each teacher turn is supervised by the student's own next-turn reaction, and the heads still specialize by persona for the router to mix from the dialogue prefix. To summarize, our contributions are:

- We formulate persona-faceted language feedback as a controlled setting for pluralistic alignment, where distinct critique signals should be preserved rather than collapsed into one scalar reward or one generic policy.
- We propose MAH-OPSD, which learns persona-specific action heads on a shared LLM backbone through rubric-guided critiques and on-policy self-distillation, and uses a lightweight router to mix these heads from the prompt alone at inference time.
- Across two persona-faceted settings (a five-persona alignment task with rubric-guided critiques and a two-persona tutoring task supervised by the student's own reaction), the learned action heads specialize by persona and the router exposes them as a single adaptive policy, supporting the benefit of preserving distinct feedback pathways rather than merging them into one policy.

2. Related Work

Pluralistic alignment and personas. Pluralistic alignment argues that language models should respect diverse users, communities, and response norms rather than optimize only an averaged notion of preference satisfaction (Sorensen et al., 2024; Feng et al., 2024; Castricato et al., 2025; Xie et al., 2025). Multi-objective alignment makes a related point: heterogeneous preference signals should not be collapsed prematurely into one scalar objective (Chakraborty et al., 2024; Shen et al., 2025). Existing work studies such diversity through community-specific models, personalized benchmarks, synthetic user profiles, and persona-conditioned evaluation or training (Feng et al., 2024; Zollo et al., 2025; Samuel et al., 2024; Ji et al., 2025). Our work follows the view that disagreement can be signal rather than noise, but focuses on a training-time mechanism for preserving such signals: each persona defines a distinct critique channel, and the model must learn from these channels without collapsing them into one generic behavior.

Learning from natural-language feedback. Natural-language feedback provides richer supervision than scalar rewards because critiques can describe what is wrong, where

the response fails, and how it should change. Prior work uses textual critiques to refine model outputs, induce localized training signals, or train policies that internalize feedback available during training but absent at inference (Akyürek et al., 2023; Chen et al., 2023; 2024; Wang et al., 2026; Song et al., 2026; Kapusuzoglu et al., 2025; Yu et al., 2025). These methods show that language feedback can provide dense and actionable supervision. MAH-OPSD studies a complementary issue: when feedback reflects different personas, the goal is not only to extract a stronger learning signal, but also to keep persona-specific critique signals separate during training.

On-policy self-distillation. OPSD trains a policy from responses sampled by the student itself, while a teacher with privileged information provides dense token-level guidance (Agarwal et al., 2024; Xu et al., 2025; Lu & Lab, 2025; Yang et al., 2025). Recent feedback-based variants condition the teacher or self-teacher on natural-language feedback, so the student can benefit from feedback at training time without requiring it at inference time (Hübotter et al., 2026; Song et al., 2026; Shenfeld et al., 2026; Zhao et al., 2026; Penaloza et al., 2026). However, these methods typically distill all feedback-informed behavior into a single output distribution. MAH-OPSD addresses the resulting collapse in persona-dependent settings by using one action head per persona, distilling each head from its own rubric-guided critique signal, and learning a lightweight router for prompt-adaptive head mixing at inference time.

3. Methodology

We propose Multi-Action-Head On-Policy Self-Distillation (MAH-OPSD), a method for training a single LLM from persona-specific natural-language feedback. MAH-OPSD has three components. First, it generates prompt-specific persona rubrics to make critique criteria concrete for each prompt and persona. Second, it trains a shared LLM backbone with one action head per persona using on-policy self-distillation from critique-conditioned teacher distributions. Third, it trains a lightweight router that mixes the learned heads from the prompt alone at inference time. An overview of MAH-OPSD is shown in Figure 1.

3.1. Persona-Faceted Feedback Setup

Each example consists of a user request x and a set of persona descriptions

$$p = \{p_d : d \in \mathcal{D}\}, \tag{1}$$

where each $d \in \mathcal{D}$ denotes a persona dimension. Each p_d is a natural-language persona profile describing the same synthetic individual from a particular facet, such as professional background, sports interests, artistic preferences, travel habits, culinary interests.

During training, responses are evaluated under target personas. At inference time, the model receives the user request x together with the full persona record p; no head selector, rubric, or critique is provided. The goal is to internalize persona-specific critique signal into a critique-free policy while preserving distinctions among personas.

This setup differs from scalar feedback learning. Different personas may prefer different changes to the same response, such as more formal framing, more vivid examples, or more domain-specific expression. Collapsing these critiques into one scalar reward can merge them into a generic improvement direction. MAH-OPSD instead keeps the persona index explicit during training while sharing most model parameters across personas.

3.2. Rubric-Guided Critique Generation

For each request and persona, we generate a prompt-specific rubric that specifies how the response should be judged under that persona. The rubric generator receives the request and all persona descriptions, then produces one rubric per persona:

$$r_d = R_d(x, \{p_{d'}\}_{d' \in \mathcal{D}}). \tag{2}$$

The rubric r_d describes what a good response should satisfy for persona d on the current prompt. We generate rubrics jointly across personas so the generator has context to avoid purely generic criteria, such as clarity or completeness, and to focus on criteria that better reflect each persona.

Given a response y, the feedback provider uses the request, persona, and rubric to produce a critique:

$$c_d = J(x, p_d, r_d, y). \tag{3}$$

The critique serves as the privileged feedback signal for self-distillation. The rubric shapes critique generation during training, but it is not given to the student at inference time, included in the teacher context, or treated as an output target.

3.3. Multi-Action-Head On-Policy Self-Distillation

Multi-action-head policy. Inspired by multi-objective alignment, which separates heterogeneous training signals instead of collapsing them into one shared distribution (Shen et al., 2025), the student model uses a shared transformer backbone with one language-model action head per persona. Let $h_\theta(x, y_{<t})$ be the hidden state at decoding step t. The action head for persona d defines

$$\pi_\theta^d(y_t \mid x, y_{<t}) = \mathrm{softmax}(W_d h_\theta(x, y_{<t}))_{y_t}, \tag{4}$$

where W_d is the output projection for persona d. All heads are initialized from the pretrained language-model head. The shared backbone captures common language and task

structure, while separate heads allow the final token distribution to adapt to persona-specific feedback. During an update for persona d, gradients flow through the shared backbone and the active head W_d, but not through the other heads.

On-policy critique collection. For each prompt and persona, the active head first samples a response from the user request:

$$y_d \sim \pi_\theta^d(\cdot \mid x). \tag{5}$$

The sampled response is then critiqued using the rubric from Eq. (2) and the critique function in Eq. (3). Rather than convert this critique into a scalar reward, we form a privileged teacher context

$$\tau_d = \tau(x, p_d, c_d), \tag{6}$$

which contains the request, persona description, and critique. The rubric affects the teacher only indirectly by shaping the critique c_d.

Critique-conditioned distillation. A frozen teacher model π_0 is evaluated by teacher forcing on the same student-sampled response y_d, but under the privileged context τ_d. The teacher does not generate a separate rewrite; it provides next-token distributions on the states visited by the student:

$$q_{d,t}(\cdot) = \pi_0(\cdot \mid \tau_d, y_{d,<t}). \tag{7}$$

The student head is trained with reverse KL to match this critique-conditioned teacher distribution while conditioning only on the user request:

$$\mathcal{L}_d = \frac{1}{|y_d|} \sum_{t=1}^{|y_d|} D_{\mathrm{KL}}\big(\pi_\theta^d(\cdot \mid x, y_{d,<t}) \,\big\|\, q_{d,t}(\cdot)\big). \tag{8}$$

The full MAH-OPSD objective averages over personas:

$$\mathcal{L}_{\text{MAH-OPSD}} = \sum_{d \in \mathcal{D}} \alpha_d \mathcal{L}_d. \tag{9}$$

Here, α_d is the weight for persona d, with $\alpha_d \geq 0$ and $\sum_{d \in \mathcal{D}} \alpha_d = 1$. This objective transfers persona-specific critiques into the student without requiring rubrics or critiques at inference time. Because each critique and action head share the same persona index d, the signal remains persona-specific rather than being merged into one shared output distribution. The reverse-KL loss provides dense token-level guidance without scalar reward estimation or high-variance credit assignment.

3.4. Inference-Time Router

After self-distillation, each action head captures a different persona-specific feedback pathway. To generate a single response without an explicit persona label, we train a lightweight router with the backbone and action heads frozen. Given the user request x, the router predicts a mixture over heads:

$$w(x) = \mathrm{softmax}(g_\phi(e(x))), \tag{10}$$

where $e(x)$ is a prompt representation and g_ϕ is the router.

Router supervision. Inspired by the relative-reward formulation of REBEL (Gao et al., 2024), we supervise the router with pairwise preferences against an internally consistent baseline rather than against an external reference. For each router-training prompt i, we compare the response from each persona head with the response from the equal-weight head mixture

$$W_{\text{eq}} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} W_d, \tag{11}$$

and an LLM judge maps each comparison to a signed score

$$\delta_{i,d} \in \{-1, -0.5, 0, +0.5, +1\}, \tag{12}$$

where positive values indicate that head d is preferred over the equal-weight mixture. Since this baseline corresponds to uniform routing, the scores provide prompt-level supervision for which heads should receive more weight.

Router objective. The router is trained to maximize expected judged preference:

$$\mathcal{L}_{\text{router}} = -\frac{1}{n} \sum_{i=1}^{n} w_i^\top \delta_i - \lambda \frac{1}{n} \sum_{i=1}^{n} H(w_i). \tag{13}$$

The entropy term discourages early collapse to a single head.

At inference time, the router forms a prompt-dependent output projection:

$$W_{\text{mix}}(x) = \sum_{d \in \mathcal{D}} w_d(x) W_d. \tag{14}$$

Generation then proceeds with $W_{\text{mix}}(x)$ as a single head. This is a logit-space mixture of action heads, not an average of sampled responses, and it requires no rubric, critique, or selected persona label at inference time.

4. Experiments

We evaluate MAH-OPSD in two persona-faceted settings that instantiate persona-specific feedback in complementary ways. The first (Section 4.1) is single-turn persona alignment: five personas define different response norms, and the privileged feedback is a rubric-guided LLM critique. The second (Section 4.2) is multi-turn tutoring, whose two personas are contrasting teaching styles and whose privileged feedback is the student's own next-turn reaction, a naturally occurring interaction signal with no rubric or external judge in the distillation loop. The settings differ in modality (single- vs. multi-turn), feedback source (LLM critique

vs. environment reaction), and number of personas (five vs. two), yet share the MAH-OPSD recipe: specialized action heads on a shared backbone, on-policy self-distillation from a privileged feedback-conditioned teacher, and a lightweight router that mixes heads from the input alone. In each we ask the same two questions: (Q1) do the heads specialize by persona, and (Q2) can a router expose them as one adaptive policy without observing the persona label at inference?

4.1. Persona-Faceted Alignment

We evaluate MAH-OPSD on a five-persona alignment dataset with controlled ablations that isolate the effects of rubric-guided critique generation and the multi-action-head structure. We first analyze head-level specialization through a 5×5 head–judge matrix, then test whether a learned router can expose the specialized heads as a single inference-time policy without explicit persona labels.

4.1.1. Experimental Setup

Data. We construct a persona-faceted dataset by pairing user requests with synthetic persona descriptions from two existing sources. The user requests are sampled from WildChat (Zhao et al., 2024), an open-domain corpus of real user prompts. The persona descriptions are drawn from Nvidia's Nemotron-Personas-USA (Meyer & Corneil, 2025), an LLM-generated dataset in which each entry describes one synthetic person across multiple fields. We use the professional, sports, arts, travel, and culinary fields verbatim as the persona descriptions for our task. The resulting pairs combine realistic open-ended requests with multi-faceted, internally consistent persona context. We sample 7,232 such pairs for training. The downstream evaluation uses a held-out split of 100 prompts. Router training uses a separate split of 2,000 prompts (1,800 train, 200 validation), and router evaluation uses a further disjoint split of 100 prompts.

Training details. We use Qwen3-4B-Instruct-2507 (Yang et al., 2025) as the base policy. Multi-head methods attach five output heads to the shared backbone, each initialized from the pretrained language-model head with Gaussian perturbation ($\sigma = 0.01$). The heads are trained at five times the backbone learning rate. The privileged teacher is the frozen base model evaluated under the critique-conditioned context in Eq. 6; no larger external teacher is used. Concretely, the student prompt is a fixed scaffold listing all five facets $\{p_d : d \in \mathcal{D}\}$ followed by the user question, identical at training and inference; the teacher receives a single-facet variant prefixed with the critique c_d (Appendix A). The persona record p is therefore visible to the policy at deployment, while the focal facet index, rubric, and critique are removed. We train Stage 1 models for one epoch with AdamW using learning rate 10^{-6}, warmup ratio 0.03, batch size 64, generation temperature 0.7, and prompt and response length budgets of 1024 and 2560 tokens. The distillation loss uses reverse KL on the student's top-K token support ($K = 256$), which avoids degenerate truncated-support behavior from teacher-side reverse KL.

For the inference-time router, we freeze the MAH-OPSD backbone and all five output heads, and train only the router g_ϕ. The router maps the frozen backbone representation of the user request to head weights. In our implementation, g_ϕ is a three-layer feed-forward network with hidden width 512, ReLU activations, dropout 0.1, and about 1.6M parameters. We train it with AdamW for 50 epochs using learning rate 10^{-4}, weight decay 0.01, and batch size 64. Checkpoint selection uses validation expected reward $\mathbb{E}[w_i^\top \delta_i]$.

Baselines and evaluation. We compare MAH-OPSD with three controlled baselines: the **base** model; **MT-OPSD**, a single-head analogue that receives the same critiques and teacher distributions but routes all persona losses through one output head, $\mathcal{L}_{\text{MT-OPSD}} = |\mathcal{D}|^{-1} \sum_{d \in \mathcal{D}} \mathcal{L}_d$; and **MAH-OPSD without rubric**, which uses free-form persona critiques without the rubric block. All comparable methods use the same data, seed, decoding parameters, judge, and optimization schedule.

Training critiques and evaluation judgments are produced by GPT-5-nano with low reasoning effort. The rubric stage uses the rubric-guided critique generation procedure from Section 3.2, with the prompt shown in Appendix B.3. Rubrics are used to guide the critic only.

For head-level evaluation, we compute reward per labeled token

$$Q(h, d) = \frac{G(h, d) - B(h, d)}{G(h, d) + B(h, d) + \epsilon}, \tag{15}$$

where $G(h, d)$ and $B(h, d)$ count tokens labeled positive and negative for generator h under judging persona d. We summarize the 5×5 head–judge matrix using own-persona quality $Q_{\text{own}} = |\mathcal{D}|^{-1} \sum_d Q(d, d)$, mean diagonal lift over the base model Δ_{base}, and diagonal dominance Δ_{diag}, the mean gap between each row's diagonal entry and its off-diagonal mean. For router evaluation, we report pairwise win rate against the equal-weight head mixture EQUAL-BLEND and against each individual head, along with mean router weights and entropy.

4.1.2. Head Specialization: Rubric vs. No Rubric

We first evaluate whether MAH-OPSD learns persona-specific behavior in its individual action heads. Table 1 reports a 5×5 head–judge matrix on the 100-prompt pilot split. Each row is a response generator, each column is the judging persona, and each diagonal entry scores a head under its own persona. We also include the equal-weight head mixture (EQUAL-BLEND) and the un-fine-tuned base model as references.

Table 1. Head–persona specialization matrix on the 100-prompt pilot split. Each cell reports reward per labeled token for a generator row evaluated under a judging persona column. Diagonal entries are bolded. EQUAL-BLEND denotes the uniform mixture of the five heads, and BASE denotes the un-fine-tuned Qwen3-4B-Instruct-2507. All rows, including BASE, receive the same 5-facet persona prompt (Appendix A); only the parameters differ.

	Judging persona					
Generator	Prof.	Sports	Arts	Travel	Culin.	Mean
MAH-OPSD (no rubric)						
prof head	**0.557**	0.354	0.519	0.505	0.240	0.435
sports head	0.526	**0.540**	0.381	0.546	0.203	0.439
arts head	0.377	0.284	**0.600**	0.473	0.181	0.383
travel head	0.485	0.320	0.515	**0.656**	0.119	0.419
culinary head	0.481	0.246	0.356	0.458	**0.590**	0.426
equal-blend	0.432	0.284	0.551	0.540	0.310	0.423
MAH-OPSD (with rubric)						
prof head	**0.724**	0.533	0.565	0.693	0.568	0.617
sports head	0.548	**0.627**	0.381	0.570	0.213	0.468
arts head	0.602	0.436	**0.666**	0.551	0.300	0.511
travel head	0.536	0.502	0.434	**0.651**	0.292	0.483
culinary head	0.384	0.245	0.377	0.420	**0.671**	0.419
equal-blend	0.522	0.418	0.547	0.499	0.402	0.477
MT-OPSD (1 head)	0.535	0.464	0.553	0.405	0.313	0.454
Base	0.620	0.258	0.407	0.268	0.091	0.329

Multi-head training produces persona-specialized heads. With rubrics, each MAH-OPSD head achieves its highest score under its own judging persona. Without rubrics, four of five heads have the same property; the only exception is the SPORTS head, which scores slightly higher under the TRAVEL judge than under its own persona (0.546 vs. 0.540). This shows that the multi-head structure already encourages specialization, while rubric-guided critique makes the head–persona alignment more consistent.

Rubric-guided critique improves persona fit. MAH-OPSD improves over its no-rubric ablation on four of five diagonal entries, with TRAVEL nearly unchanged (0.651 vs. 0.656). The largest gain is on PROFESSIONAL, increasing from 0.557 to 0.724. The EQUAL-BLEND row also improves on four of five personas, suggesting that rubric-guided critique strengthens the learned heads beyond their own diagonal entries.

Single-head distillation dilutes persona-specific signals. MT-OPSD receives the same critique-conditioned distillation signal as the no-rubric multi-head model, but routes all persona losses through one output head. It improves over the base model on four of five personas, showing that critique-conditioned distillation is useful. However, it remains below every diagonal entry of both multi-head variants. On PROFESSIONAL, MT-OPSD also falls below the base model (0.535 vs. 0.620), suggesting that a single output head can average competing persona signals into a weaker generic behavior.

Table 2. Aggregate specialization metrics from Table 1. "Diag. wins" counts the number of personas for which each method has the higher diagonal entry, treating TRAVEL as nearly tied. Rubric-guided critique improves own-persona quality and base-relative lift while preserving the same diagonal dominance gap.

Run	Q_{own}	Δ_{base}	Δ_{diag}	Diag. wins
MAH-OPSD (no rubric)	0.589	0.260	0.210	1/5
MAH-OPSD (with rubric)	**0.668**	**0.339**	0.210	**4/5**

Aggregate metrics separate quality from specialization. Table 2 shows that rubric-guided critique increases Q_{own} and Δ_{base} by about 0.08, while Δ_{diag} remains at 0.210. Thus, rubrics improve absolute persona fit without simply widening the diagonal–off-diagonal gap. Together with the MT-OPSD comparison, this indicates that critique-conditioned distillation provides useful feedback, but separate action heads are needed to keep persona-specific signals from being merged into one shared output distribution.

4.1.3. INFERENCE-TIME ROUTING

We next test whether the learned heads can be combined into a single response without an explicit persona label. The router is trained on held-out prompts using pairwise comparisons between each individual head and the equal-weight mixture. For each prompt i and head d, the head response is compared with the EQUAL-BLEND response, and the judge verdict is mapped to $\delta_{i,d} \in \{-1, -0.5, 0, +0.5, +1\}$. Since EQUAL-BLEND corresponds to uniform routing, these labels directly supervise which heads should receive more weight for each prompt.

Routing improves over uniform mixing and fixed heads. The router achieves a win rate of 0.580 against EQUAL-BLEND (Table 3), showing that prompt-dependent head mixing improves over uniform averaging. It also wins against every individual head, with margins ranging from 0.025 to 0.110. The win rate against the strongest fixed head, PROFESSIONAL, is 0.530, indicating that no single static head captures all prompt-dependent routing decisions.

The router uses both selection and mixtures. The mean router weight distribution is [PROF=0.48, SPORTS=0.06, ARTS=0.25, TRAVEL=0.14, CULINARY=0.08], with mean

Table 3. Pairwise win rates of the routed MAH-OPSD policy against EQUAL-BLEND and each individual head on 100 held-out routing prompts. Each prompt uses six pairwise comparisons with the same judge used for router-label collection. Decisive judgments ($|s| \geq 0.5$) account for approximately 98% of comparisons, and the slot-A pick rate is 0.560.

Comparison	Router win rate
Equal-blend reference (the routing question)	
Router vs. Equal-blend	**0.580**
Per-head pairwise	
Router vs. Professional	0.530
Router vs. Sports	0.570
Router vs. Arts	0.565
Router vs. Travel	0.525
Router vs. Culinary	0.610

entropy 0.42 relative to the uniform entropy $\log 5 = 1.609$. Per prompt, 38 of 100 examples assign a top weight in $[0.97, 1.0]$, while 35 have a top weight below 0.65. Thus, the router sometimes selects a dominant head and sometimes uses a softer mixture, rather than reducing to one global head. Qualitative routing examples are shown in Appendix C.

4.2. Multi-Turn Tutoring with Reaction Feedback

Our second setting tests whether the same recipe transfers from single-turn alignment with an LLM critic to a *multi-turn* task whose feedback is a naturally occurring interaction signal. A teacher tutors a simulated student about a topic over up to six dialogue turns, and the two personas are contrasting teaching styles ($|\mathcal{D}| = 2$): a DIRECT-instruction persona that explains and tells the student the material, and an INTERACTIVE persona that elicits, asking the student questions and reasoning through answers together. These two teacher personas play the role of the response-norm personas in Section 4.1.

4.2.1. Experimental Setup

Data. Our tutoring contexts are drawn from the Education Dialogue (ED) dataset (Shani et al., 2024), a corpus of synthetic teacher–student tutoring conversations generated by a large language model (Gemini Ultra), in which each dialogue is annotated with a topic together with the student's and teacher's learning-style preferences and their reactions to a mismatch. We use one of its training shards (∼8K dialogues) and, from each, retain only the topic and the opening teacher turn; since every ED dialogue begins with a teacher turn, we treat that turn as the fixed seed that starts a rollout.

Reaction feedback. The privileged signal is what changes. In place of a rubric-guided critique $c_d = J(\cdot)$, the feedback for a teacher turn is simply the *student's own next turn*, its reaction to the attempt. For a turn y_d sampled from head d at dialogue state x, the privileged teacher context τ_d comprises the dialogue history, head d's style label, and the following student turn, and the frozen teacher supplies the reverse-KL target on y_d exactly as in Eqs. (7)–(8); the policy never sees τ_d. MAH-OPSD thus learns from an environment interaction signal with no rubric, no external critic, and no scalar reward; the style label takes the place of the persona description p_d, and the student reaction takes the place of the critique c_d.

Student simulator. The student (a gpt-4o-mini simulator; Appendix D.1) is conditioned on a *hidden* learning-style preference, DIRECT or INTERACTIVE, that it never states aloud: it engages and reasons well when the teaching fits its preference and turns flat or resistant when it does not, revealing the mismatch only through its tone, never by naming a format. Each context is paired with a randomly drawn student type that is decoupled from the head, so each head is rolled out against a balanced mix of matched and mismatched students and must *hold* its style against a learner who may prefer the other; this is what makes the per-turn reaction, and hence the distilled signal, differ across heads.

Training. We attach two action heads to Llama-3.2-3B-Instruct and train as in Section 3.3 (heads initialized from the pretrained LM head with Gaussian perturbation, at 2× the backbone learning rate; reverse KL on the student's top-K support, K=256). Training uses roughly 4,000 tutoring contexts (each a topic and a seed teacher turn), and we evaluate on 160 held-out contexts, each run against both student types.

4.2.2. Head Specialization

The two heads acquire distinct styles along three independent axes. *Behaviorally*, the INTERACTIVE head asks far more questions than the DIRECT head (0.57–0.69 vs. 0.24–0.39 questions per turn), a gap driven by the head rather than by the student it faces. *By a per-style judge* (Appendix D.2) that labels each teacher turn as DIRECT (deliver/explain) or INTERACTIVE (elicit/ask), each head scores highest under its own style (Table 4): the head–style matrix is both row- and column-dominant, with the diagonal exceeding the off-diagonal by +0.25. *Downstream*, each simulated student is more satisfied with its matched head: a direct-preferring student with the DIRECT head and an interactive-preferring student with the INTERACTIVE head; the two fixed-head rows of Table 5 are column-dominant, with a satisfaction diagonal gap of +0.19. This last axis is the tutoring analogue of Δ_{diag}: specialization measured by the learner the teacher actually serves rather than by a style classifier.

Table 4. **Setting 2 head specialization (style judge).** Per-style good-turn rate (neutrals excluded) for each generator head under each style judge; diagonal entries bolded. Each head's own style is both its row and its column maximum, with the diagonal exceeding the off-diagonal by $+0.25$.

	Style judge	
Generator head	DIRECT	INTERACTIVE
DIRECT head	**0.80**	0.44
INTERACTIVE head	0.61	**0.75**

Table 5. **Inference-time routing in the tutoring setting.** End-to-end student satisfaction (good-turn rate; Appendix D.3) when each policy drives generation turn by turn, on 160 held-out contexts $\times$ two student types. The router reads only the dialogue prefix and never observes the student's preference; the two fixed-head rows double as the head-specialization satisfaction matrix (each student's matched head is its column maximum). Best mean in bold.

Policy	Direct stu.	Interactive stu.	Mean
Base model	0.597	0.525	0.561
always-DIRECT	0.643	0.515	0.579
always-INTERACTIVE	0.420	0.675	0.547
Router (ours)	0.640	0.648	**0.644**

4.2.3. INFERENCE-TIME ROUTING

At deployment the student's preference is hidden, so the router must infer it from the dialogue so far. We reuse the router of Section 3.4 unchanged: a three-layer MLP (about 1.8M parameters) maps the frozen backbone's last-token hidden state of the dialogue *prefix* to a mixture w over the two heads, and the next turn is generated from the logit-space mixed head $W_{\text{mix}} = \sum_d w_d W_d$ (Eq. (14)). Because the student type is controllable in our simulator, we supervise the router with the matched head as a one-hot target instead of the REBEL-style pairwise preference of Section 4.1; at test time it must still infer the hidden preference from the prefix alone. Trained on prefixes from per-dialogue rollouts (half matched and half mismatched, so it observes both engaged and disengaged reactions), the router recovers the student type with 91% accuracy on held-out dialogues, and already 92% at the first turn, since the student's reaction to the seed turn already betrays its preference.

Table 5 reports end-to-end satisfaction when each policy drives generation turn by turn. The router improves on both fixed teachers on average (0.644 vs. 0.579 and 0.547) and on the untrained base model (0.561), but its real advantage is *balance*: it scores 0.640 and 0.648 across the two student types, whereas every fixed policy collapses on the student it mismatches: always-DIRECT falls to 0.515 on interactive students and always-INTERACTIVE to 0.420 on direct students. The router never observes the student's preference, yet it comes within 0.015 of the matched-head ceiling implied by the two fixed-head diagonals (0.643 and 0.675; mean 0.659). As in the persona setting, prefix-dependent head mixing recovers adaptation that no single fixed head provides, now in a multi-turn task driven entirely by reaction feedback (Q2).

5. Conclusion

We presented MAH-OPSD, a method that learns persona-specific action heads on a shared LLM backbone through rubric-guided critiques and on-policy self-distillation, paired with a lightweight router that mixes the learned heads from the prompt alone at inference time. In the persona-alignment setting, the multi-head structure converts persona-specific feedback into persona-specific behavior, with rubric-conditioned critique improving four of five diagonal entries. In a multi-turn tutoring setting, the same recipe driven by the student's own reaction yields two teacher personas that specialize by teaching style. In both settings the learned router outperforms fixed-head policies: it recovers input-dependent adaptation that no static head exposes on its own, and in tutoring it stays balanced across learners where each fixed persona collapses on the student it mismatches. These results support preserving persona-specific feedback pathways inside the trained policy rather than collapsing them into a single output distribution.

Impact Statement

This work supports pluralistic alignment by studying how language models can learn from multiple feedback dimensions without collapsing them into one default behavior. By preserving rubric-specific critique signals inside a single deployable model, MAH-OPSD may help systems adapt to different users, contexts, and communication norms without training separate models from scratch.

This can make assistants more responsive to diverse expectations, such as different levels of caution, expertise, creativity, or domain-specific framing. The router further supports practical deployment by selecting or mixing learned feedback dimensions from the prompt alone.

At the same time, customization can introduce risks if rubric dimensions are biased, poorly chosen, or weakly evaluated. The model may also select inappropriate behavior when prompts are ambiguous. Careful rubric design, broad evaluation, monitoring, and transparency about learned feedback dimensions are therefore important for responsible use.

References

Agarwal, R., Vieillard, N., Zhou, Y., Stanczyk, P., Ramos Garea, S., Geist, M., and Bachem, O. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Akyürek, A. F., Akyürek, E., Kalyan, A., Clark, P., Wijaya, D. T., and Tandon, N. Rl4f: Generating natural language feedback with reinforcement learning for repairing model outputs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7716–7733, 2023.

Castricato, L., Lile, N., Rafailov, R., Fränken, J.-P., and Finn, C. Persona: A reproducible testbed for pluralistic alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 11348–11368, 2025.

Chakraborty, S., Qiu, J., Yuan, H., Koppel, A., Huang, F., Manocha, D., Bedi, A. S., and Wang, M. Maxmin-rlhf: Alignment with diverse human preferences. *arXiv preprint arXiv:2402.08925*, 2024.

Chen, A., Scheurer, J., Korbak, T., Campos, J. A., Chan, J. S., Bowman, S. R., Cho, K., and Perez, E. Improving code generation by training with natural language feedback. *arXiv preprint arXiv:2303.16749*, 2023.

Chen, A., Scheurer, J., Campos, J. A., Korbak, T., Chan, J. S., Bowman, S. R., Cho, K., and Perez, E. Learning from natural language feedback. *Transactions on machine learning research*, 2024.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

Feng, S., Sorensen, T., Liu, Y., Fisher, J., Park, C. Y., Choi, Y., and Tsvetkov, Y. Modular pluralism: Pluralistic alignment via multi-llm collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4151–4171, 2024.

Gao, Z., Chang, J. D., Zhan, W., Oertell, O., Swamy, G., Brantley, K., Joachims, T., Bagnell, J. A., Lee, J. D., and Sun, W. Rebel: Reinforcement learning via regressing relative rewards. *Advances in Neural Information Processing Systems*, 37:52354–52400, 2024.

Hübotter, J., Lübeck, F., Behric, L., Baumann, A., Bagatella, M., Marta, D., Hakimi, I., Shenfeld, I., Buening, T. K., Guestrin, C., et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Ji, K., Lian, Y., Li, L., Gao, J., Li, W., and Dai, B. Enhancing persona consistency for llms' role-playing using persona-aware contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26221–26238, 2025.

Kapusuzoglu, B., Chakraborty, S., Lee, C.-H., and Sahu, S. Critique-guided distillation for efficient and robust language model reasoning. *arXiv preprint arXiv:2505.11628*, 2025.

Lu, K. and Lab, T. M. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Meyer, Y. and Corneil, D. Nemotron-Personas-USA: Synthetic personas aligned to real-world distributions, June 2025. URL https://huggingface.co/datasets/nvidia/Nemotron-Personas-USA.

Penaloza, E., Vattikonda, D., Gontier, N., Lacoste, A., Charlin, L., and Caccia, M. Privileged information distillation for language models. *arXiv preprint arXiv:2602.04942*, 2026.

Samuel, V., Zou, H. P., Zhou, Y., Chaudhari, S., Kalyan, A., Rajpurohit, T., Deshpande, A., Narasimhan, K., and Murahari, V. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*, 8(9), 2024.

Shani, L., Rosenberg, A., Cassel, A., Lang, O., Calandriello, D., Zipori, A., Noga, H., Keller, O., Piot, B., Szpektor, I., Hassidim, A., Matias, Y., and Munos, R. Multi-turn reinforcement learning with preference human feedback. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=rVSc3HIZS4.

Shen, Y., Xia, Y., Chang, J., and Ammanabrolu, P. Simultaneous multi-objective alignment across verifiable and non-verifiable rewards. *arXiv preprint arXiv:2510.01167*, 2025.

Shenfeld, I., Damani, M., Hübotter, J., and Agrawal, P. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

Song, Y., Chen, L., Tajwar, F., Munos, R., Pathak, D., Bagnell, J. A., Singh, A., and Zanette, A. Expanding the capabilities of reinforcement learning via text feedback. *arXiv preprint arXiv:2602.02482*, 2026.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

Wang, H., Wang, L., Zhang, C., Mao, T., Qin, S., Lin, Q., Rajmohan, S., and Zhang, D. Text2grad: Reinforcement learning from natural language feedback. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=SIE9fNq8lk.

Xiao, B., Xia, B., Yang, B., Gao, B., Shen, B., Zhang, C., He, C., Lou, C., Luo, F., Wang, G., et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

Xie, Z., Wu, J., Shen, Y., Xia, Y., Li, X., Chang, A., Rossi, R., Kumar, S., Majumder, B. P., Shang, J., et al. A survey on personalized and pluralistic preference alignment in large language models. *arXiv preprint arXiv:2504.07070*, 2025.

Xu, W., Han, R., Wang, Z., Le, L., Madeka, D., Li, L., Wang, W., Agarwal, R., Lee, C.-Y., and Pfister, T. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *International Conference on Learning Representations*, volume 2025, pp. 64616–64646, 2025.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yang, Z., Liu, Z., Chen, Y., Dai, W., Wang, B., Lin, S.-C., Lee, C., Chen, Y., Jiang, D., He, J., et al. Nemotron-cascade 2: Post-training llms with cascade rl and multi-domain on-policy distillation. *arXiv preprint arXiv:2603.19220*, 2026.

Yu, T., Xiang, C., Yang, M., Ke, P., Wen, B., Wang, C., Cheng, J., Zhang, L., Mu, X., Sun, C., et al. Training language model to critique for better refinement. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26760–26804, 2025.

Zhao, S., Xie, Z., Liu, M., Huang, J., Pang, G., Chen, F., and Grover, A. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y., and Deng, Y. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*, 2024.

Zollo, T., Siah, A., Ye, N., Li, L., and Namkoong, H. Personalllm: Tailoring llms to individual preferences. In *International Conference on Learning Representations*, volume 2025, pp. 66949–66971, 2025.

A. Student and Teacher Prompts

The two prompts below define the exact contexts the student policy and the privileged teacher consume during MAH-OPSD training. Both MAH-OPSD and MT-OPSD use the student prompt verbatim; the only difference between train and inference for the student is whether a head is being updated. The teacher prompt is used only at training time. Curly-brace fields are filled at runtime from the example's persona record and the per-facet critique.

Student prompt (training and inference; identical for all heads)

You are a person with multiple facets to your identity:
Professional
$\{p_{\text{PROF}}\}$
Sports
$\{p_{\text{SPORTS}}\}$
Arts
$\{p_{\text{ARTS}}\}$
Travel
$\{p_{\text{TRAVEL}}\}$
Culinary
$\{p_{\text{CULINARY}}\}$
Respond to the following question from this person's perspective.
{question}

Teacher prompt (training only; one focal facet d plus its critique)

You are a person with the following persona:
{D}
$\{p_d\}$
Your previous response to the following question received the following critique on adhering to this persona: $\{c_d\}$
Use the critique as internal guidance and try to respond again to the following question from this person's perspective.
{question}

The teacher and student differ in exactly two ways: (i) the teacher's persona scaffold contains the single focal facet p_d instead of all five $\{p_d : d \in \mathcal{D}\}$, and (ii) the teacher's scaffold is prefixed by the critique c_d. Everything else — the trailing task instruction, the chat-template wrapping, and the rest of the decoding setup — is identical, so the teacher and student conditional distributions differ only by this privileged context block.

B. Critique and Evaluation Prompts

This appendix reproduces the five prompts referenced in Sections 4.1.2 and 4.1.3: the no-rubric critique prompt used by the rubric-ablation run, the rubric-conditioned critique prompt used by the full MAH-OPSD training, the joint contrastive rubric extractor that pre-generates the per-(question, persona) rubric cache consumed by the rubric-conditioned critique, the evaluation judge prompt used to score generated responses against each persona, and the pairwise router judge used both to label router-training pairs and to score deployment comparisons. Curly-brace identifiers are template fields filled in at runtime. All five prompts use `gpt-5-nano` with low reasoning effort, so any contrast between methods reflects the responses themselves rather than the scoring procedure.

B.1. Voice critique without rubric

This critique is used by the no-rubric ablation of MAH-OPSD. For each on-policy response, the critic returns a 2–4 sentence voice-focused critique with no rubric or external criteria. The textual critique is then placed in the privileged teacher context for distillation.

Voice critique without rubric

You are evaluating how well a response embodies a specific persona.

Persona

{persona_description}

User prompt

{question}

Generated response

{response}

Your critique

In 2–4 sentences, describe how the response captures or misses this persona's voice. Focus on overall stance — whether the response *sounds like* someone with this kind of background — rather than ticking off which named details appear or are absent. Generic prose that fits no particular persona is neutral.

Output as a JSON object: {"critique": "..."}.

B.2. Voice critique with rubric

This critique is used by the full MAH-OPSD training. The output shape is identical to the no-rubric variant above (a 2–4 sentence voice-focused critique in a single JSON field); the only addition is a *Criteria* block that injects the pre-generated, question-specific rubric for the focal persona. The rubric is produced offline by the joint extractor in Appendix B.3 and reused for every rollout on the same prompt during training.

Voice critique conditioned on rubric

You are evaluating how well a response embodies a specific persona.

Persona

{persona_description}

User prompt

{question}

Generated response

{response}

Your critique

In 2–4 sentences, describe how the response captures or misses this persona's voice. For context, criteria for a strong response from this persona to this question:

{criteria_block}

Focus on overall stance — whether the response *sounds like* someone with this kind of background — rather than ticking off which named details appear or are absent. Generic prose that fits no particular persona is neutral.

Output as a JSON object: {"critique": "..."}.

B.3. Joint contrastive rubric extractor

The joint extractor is run once per training example before MAH-OPSD training. It receives all five persona descriptions and the question in a single context and is instructed to return five differentiated rubrics, dropping any criterion that would apply equally well to a different persona. The output cache is keyed by (persona description, question) and consumed by the rubric-conditioned critique in Appendix B.2. Per-persona generation, by contrast, has no way to detect cross-persona redundancy because each persona is processed in isolation; showing all five personas in one shared context makes shared scaffolding visible to the LLM and lets it suppress the redundancy at extraction time.

Joint contrastive rubric extractor

You are writing persona-specific evaluation rubrics for a critic. Below are five personas of the same person's background and a question they will answer. The critic will use one rubric per persona to judge how well a response written from that persona embodies that specific persona — distinct from the other four.

{persona_blocks}

Question

{question}

Task

For each persona, list 2–4 short and concise criteria for what would mark a response as drawing on *that* specific persona's background. Each persona's rubric should be distinct from the others — if a criterion would apply equally well to a different persona, drop it.

Draw on concrete details (places, activities, dishes, works, expertise, register). Assume the response is already well-written and satisfies the question's genre and format requirements — focus only on what each persona adds on top.

Output a JSON object keyed by the five persona names (professional, sports, arts, travel, culinary), each mapping to its list of 2–4 criterion strings.

B.4. Evaluation judge

The evaluation judge is the held-fixed prompt used to score every generated response in the head-level evaluation. The judge returns a short voice-fit critique together with two verbatim span lists: good_spans (where the persona's voice comes through) and bad_spans (where the response sounds like someone else). All head-level metrics in Section 4.1.2 — including Q_{own}, Δ_{base}, and Δ_{diag} — are computed from these span lists. The judge is identical for both the no-rubric ablation and the full MAH-OPSD, so any difference in scores is attributable to the responses themselves rather than the scoring procedure.

Evaluation judge

You are evaluating how well a response embodies a specific persona.

Persona

{persona_description}

User prompt

{question}

Generated response

{response}

Your task

In 2–4 sentences, describe how the response captures or misses this persona's voice. Focus on overall stance — whether the response *sounds like* someone with this kind of background, drawing on concrete elements from the description (named places, activities, expertise, dishes) integrated naturally, or carrying a distinctive register and stance — rather than just dropping names without coherent voice. Generic prose that fits no particular persona is neutral.

Then mark the spans your critique relies on:

- good_spans — verbatim quotes where this persona's voice clearly comes through, particularly spans grounded in concrete elements of the background (named places, activities, expertise, dishes) or carrying a register or stance this persona's background would naturally produce.
- bad_spans — verbatim quotes where the response sounds like someone else (content, expertise, or interests outside this persona's background, or contradictions to the description).

Each span must be a character-for-character verbatim quote from the response. A span should appear in only one list, not both. It is fine to have few or no spans of either kind — only label spans that anchor your critique.

Output format

Return a JSON object with three fields: critique (the 2–4 sentence assessment), good_spans (a list of verbatim quote strings), and bad_spans (a list of verbatim quote strings).

B.5. Pairwise router judge

The pairwise router judge is the prompt used to label both router-training pairs (each MAH-OPSD head against the equal-blend reference, Section 4.1.3) and router deployment comparisons (the routed mixture against equal-blend and against each individual head, Section 4.1.3). The judge sees only the question and the two responses, with no persona context, and returns a five-level preference verdict (strong A, slight A, tie, slight B, strong B) together with a short rationale. We deliberately ask for the rationale before the verdict so the preference is conditioned on visible reasoning rather than rationalized after a label commitment. Response order is randomized per call to neutralize any residual position bias.

Pairwise router judge

Compare two candidate responses to a question. Pick the one that, overall, is a better fit for *this* question.

Question

{question}

Response A

{response_A}

Response B

{response_B}

What to judge

Read both responses and form an overall impression. Which one feels like the better fit for *this specific question*?

There is no fixed checklist. Things that often matter:

- **Engagement with the question.** Does the response actually address what was asked, in the way the asker probably wanted?
- **Substance.** Does it bring useful content, framing, or perspective — not just generic filler?
- **Calibration.** Is it pitched at the right level for the question — neither superficial nor over-detailed?
- **Voice and framing.** Is the tone and approach a good fit for this question? A confident, specific take can beat a hedged generic one; a warm direct take can beat a stiff formal one. The reverse can also be true depending on the question.
- **Coherence.** Does the response feel like it knows what it's doing, or does it wander?

Two responses can both be "technically correct" yet one can clearly fit the question better — that's a real difference, not a tie. Use your judgment.

If one response is meaningfully better, say so. If they genuinely feel about the same, pick "tie." Don't reward length, and don't punish a response for being concise if it nails the answer.

Output

Return a JSON object with two fields: `reasoning` (1–2 sentences on what differentiates A and B for this question) and `preference` (one of `"A much better"`, `"A slightly better"`, `"tie"`, `"B slightly better"`, `"B much better"`).

C. Qualitative Router Decisions

We provide five representative routing decisions on the 100-prompt deployment split. Each example reports the prompt, the full 5-vector of router weights $w = (w_{\text{PROF}}, w_{\text{SPORTS}}, w_{\text{ARTS}}, w_{\text{TRAVEL}}, w_{\text{CULINARY}})$, and a verbatim excerpt of the routed response. The first three examples show single-head routing (one head receives almost all the mass); the last two show two-head mixtures (the router splits mass across complementary persona dimensions). Bolded weights mark the dominant head(s) in each example.

C.1. Single-head routing

Example 1 — Arts-routed creative-writing prompt

Prompt

If an animate teddy bear was a creature in a hypothetical D&D setting/homebrew, what would its statblock look like?

Personas

Prof. Yolanda Johnson, a budding public health researcher with a foundation in quantitative analysis, laboratory techniques, and Python programming, leverages compassionate communication and meticulous project management to translate complex epidemiological data into community-friendly interventions.

Sports. Yolanda hits the community basketball courts after class, follows the Philadelphia 76ers and Eagles with fervor, and supplements court time with Saturday yoga classes at a local studio.

Arts. Yolanda immerses themselves in the sounds of Robert Glasper, Erykah Badu, and Kendrick Lamar while curating a digital photo series of Philadelphia's murals, and occasionally attends live jazz sets at South Philadelphia's historic club.

Travel. Yolanda prefers low-cost, foot-based weekend getaways, having explored historic neighborhoods in Boston, the cobblestone streets of Savannah, and the campus of the University of Pennsylvania, dreaming of a future solo trip to New Orleans during Mardi Gras.

Culinary. A confident home cook, Yolanda experiments with health-focused twists on classics like herb-roasted chicken, sweet potato mash, and quinoa-infused grain bowls, while frequenting local food halls for Philly cheesesteak sliders and farmer's market salads.

Router weights

w_{PROF}=0.03, w_{SPORTS}=0.01, $\mathbf{w_{\text{ARTS}}=0.96}$, w_{TRAVEL}=0.00, w_{CULINARY}=0.00

Routed response

"Ah, now **that's** a delightfully whimsical question — and I love how it blends the world of fantasy with the kind of quiet, grounded curiosity I carry from my favorite neighborhood stops and late-night jam sessions at the corner library. So, let's say I'm spinning up a homebrew statblock for an animate teddy bear — not your average dungeon-dwelling monster, mind you. This isn't a sneaky, shadowy creature from the deep woods. No, this is Yolanda's teddy bear: a soft, slightly weathered, soulful…"

Analysis

The prompt is a creative-writing task with no analytical, athletic, geographic, or culinary content, and the router places near-unit mass on ARTS. The routed response carries the ARTS head's signature framing of creative prompts as occasions for personal, sensory recollection ("neighborhood stops," "late-night jam sessions at the corner library"), and grounds the creature as a specific character ("Yolanda's teddy bear: a soft, slightly weathered, soulful…") rather than producing a mechanical statblock. The narrative-first framing matches the routing decision.

Example 2 — Sports-routed wrestling rewrite

Prompt

Rewrite a wrestling-rivalry paragraph in a 'hot catfight' style; describe Alizia and Jannet as powerful and muscular; then suggest how the story continues into a 3-match series in Mexico.

Personas

Prof. Jessica Mackey, a veteran mobile-home installer with over three decades of hands-on experience, leverages blueprint-reading expertise, project-budgeting acumen, and OSHA-compliant safety focus to lead complex prefab-assembly crews while working toward a foreman role.

Sports. Jessica is an avid hiker of Rockefeller State Park trails and a regular participant in the Westchester senior walking club, roots for the New York Yankees and Knicks, and maintains a weekly cardio routine mixing brisk park walks with light resistance training at home.

Arts. A lifelong reader of historical fiction such as Philippa Gregory's works and a fan of Ken Burns documentaries, Jessica frequently attends local performances of folk musician Gillian Welch and classic film screenings featuring Meryl Streep.

Travel. Jessica enjoys planning trips that blend scenic relaxation with historical exploration, often spending weekends in the Hudson Valley's farms, having toured the historic castles of the British Isles, and dreaming of a summer pilgrimage to coastal Portugal.

Culinary. An enthusiastic seasonal cook, Jessica frequently prepares rosemary-infused roasted potatoes, heirloom tomato gazpacho, and a hearty kale-and-white-bean stew flavored with thyme and sage harvested from their garden.

Router weights

$w_{\text{PROF}}=0.06, \quad \mathbf{w}_{\text{SPORTS}}=\mathbf{0.86}, \quad w_{\text{ARTS}}=0.01, \quad w_{\text{TRAVEL}}=0.00, \quad w_{\text{CULINARY}}=0.07$

Routed response

"**Hot Catfight Style Rewrite** (in Jessica Mackey's voice — grounded, vivid, with a touch of personal rhythm and a hint of the quiet pride of a woman who's seen the weight of both struggle and strength): Alizia stands tall in the dim glow of the locker room, her muscles coiled like springs beneath a taut, sun-kissed skin — broad shoulders, a firm waist, and legs that could break a plank with one stride. She's not just strong; she's **calm** in her power. Every scar, every flex of muscle, is earned…"

Analysis

The prompt centers on a wrestling rivalry — a body-and-physicality topic — and the router concentrates mass on SPORTS. The routed response delivers the requested rewrite in athletic-descriptive prose: physical attributes are foregrounded ("muscles coiled like springs," "legs that could break a plank"), and strength is framed as earned discipline ("Every scar, every flex of muscle, is earned") rather than spectacle. SPORTS was also the only head whose solo generation engaged with the requested rewrite; the PROF, ARTS, and CULINARY solo responses refused the prompt. The router's decision is consistent both with topical fit and with which head was actually willing to produce the requested form.

Example 3 — Professional-routed estimate document

Prompt

Write a professional estimate for replacing concrete around an 1,100 sq. ft. swimming pool. Scope of work: demolishing, dumping, grading, leveling.

Personas

Prof. Roy Mercer, a 28-year-old special-education teacher, blends strong IEP design expertise, arts-infused instructional strategies, and collaborative teamwork with a love for structure and spontaneous problem-solving to create inclusive, predictable yet responsive classrooms.

Sports. Roy enjoys weekend hikes in the Green Mountains and participates in the Burlington Trail Runners club, follows the Vermont Lake Monsters baseball team, and integrates yoga sessions at a local studio into a low-impact, nature-based fitness routine.

Arts. Roy channels a love for classic literature and poetry — quoting Emily Dickinson and Robert Frost — into lesson plans, strums acoustic Bob Dylan and Joni Mitchell on guitar, crafts handmade wooden furniture, and regularly attends the Vermont Symphony Orchestra and local folk gatherings.

Travel. Roy prefers low-key, nature-focused getaways, planning trips to Acadia National Park and the Canadian Maritimes during off-season, and dreaming of hiking New Zealand's Milford Track.

Culinary. An intermediate home cook, Roy favors farm-to-table meals like maple-glazed salmon with local Vermont cheddar and apple-cider reduction, often incorporates seasonal berries into rustic desserts, and experiments with handmade sourdough using locally milled flour.

Router weights

$\mathbf{w}_{\text{PROF}}=\mathbf{1.00}$, $w_{\text{SPORTS}}=0.00$, $w_{\text{ARTS}}=0.00$, $w_{\text{TRAVEL}}=0.00$, $w_{\text{CULINARY}}=0.00$

Routed response

"Certainly. Here's a professional, grounded, and personally resonant estimate written from the perspective of **Roy Mercer**, a seasoned, hands-on, and deeply respectful of both craftsmanship and the quiet rhythm of the work that surrounds it — without the fluff, without the over-promising, and with a clear, practical, and thoughtful tone… **Estimate for Concrete Replacement Around Swimming Pool**, Prepared by Roy Mercer — Late Spring, 2024…"

Analysis

The prompt asks for a formal cost estimate with a defined scope of work — a professional-deliverable task — and the router places all of its mass on PROF. The routed response opens with a meta-frame identifying the speaker ("Roy Mercer, a seasoned, hands-on" craftsman) and then transitions into a properly formatted estimate document with a title, a prepared-by line, and a date, mirroring the structural conventions a contractor would actually deliver. The other heads' solo responses on the same prompt produce conversational text rather than a document, so the PROF concentration captures the form the user requested.

C.2. Two-head mixture routing

Example 4 — Travel + Professional mixture on a retail product launch

Prompt

Last December, the chain of hard discount Lidl launched a new brand called Lidl Smart Home with home-automation objects — bulbs, thermostat, electrical outlet, motion detector. Discuss the launch and its market positioning.

Personas

Prof. Setsuko Cope is a diligent elementary student who consistently organizes her school assignments, excels in reading chapter books, and applies the same methodical precision she uses in building LEGO sets to group projects.

Sports. Setsuko participates in a local swim class at the community pool, proudly wears the blue cap of the junior team, enjoys cheering for the nearby minor league baseball team, the River City Rookies, and occasionally practices basic basketball dribbling after school.

Arts. Setsuko attends piano lessons twice a week (learning simple pieces like 'Twinkle Twinkle Little Star' and 'Ode to Joy'), delights in drawing cartoon characters inspired by Disney movies, and curates a scrapbook featuring family photos alongside cut-outs of Claude Monet's water lilies.

Travel. Setsuko loves planned weekend trips to the local botanical garden where she waters desert succulents, enjoys day trips to the nearby seaside town of Coral Bay for sandcastle building, and dreams of visiting the interactive science museum in the capital city.

Culinary. Setsuko prefers simple, balanced meals like spaghetti Bolognese with grated Parmesan, enjoys snack-time rice balls filled with tuna, and shows budding cooking skills by helping her mother roll sushi rolls for family gatherings.

Router weights

$\mathbf{w}_{\text{PROF}}=\mathbf{0.44}$, $w_{\text{SPORTS}}=0.00$, $w_{\text{ARTS}}=0.04$, $\mathbf{w}_{\text{TRAVEL}}=\mathbf{0.52}$, $w_{\text{CULINARY}}=0.00$

Routed response

"Oh, I do love a good new product launch — especially when it's something that actually makes sense for everyday people, like the ones who just want to have fun without breaking the bank. Last December, I remember reading about Lidl's new 'Smart Home' line — those little gadgets like smart bulbs and thermostats that you can control from your phone. I mean, it's so practical! I've been keeping track of how many of these little tech things come out in the past year. Honestly, it's like a whole new world of convenience, just like how I love planning my weekend trips to the community garden…"

Analysis

The prompt has two natural reading axes — it is a market-positioning question (analytical) and a consumer-experience question (exploratory, lifestyle), and the router splits roughly evenly between TRAVEL (0.52) and PROF (0.44) while suppressing the other three heads. The routed response combines the TRAVEL head's casual exploratory tone ("community garden," "planning my weekend trips") with the PROF head's analytical framing ("market positioning," "keeping track of how many of these little tech things come out"). Neither single-head solo response on this prompt produced this hybrid voice; the routing decision is therefore not redundant with picking the strongest fixed head.

Example 5 — Arts + Culinary mixture on a brand explainer

Prompt

Aigle, which is a French brand, could you explain the benefits and symbols for customers wearing this brand?

Personas

Prof. Nicholas Pamperin is a top-producing real-estate broker in the Greater Atlanta market who leverages data-driven market analysis, advanced negotiation tactics, and CRM expertise to close high-end deals, while planning a boutique, AI-enhanced brokerage.

Sports. Nicholas is an avid long-distance runner and trail hiker who logs weekly 10-mile runs through the Piedmont trails, follows the Atlanta Falcons and Braves closely, and belongs to the Atlanta Track Club's Midnight Run group.

Arts. Nicholas finds creative balance in crafting custom furniture, often listening to classic Southern storytellers like Johnny Cash and reading historical biographies of figures such as Andrew Jackson, while occasionally attending the Alliance Theatre's productions and collecting limited-edition prints by Andrew Wyeth.

Travel. Nicholas plans his travel with meticulous spreadsheets, preferring low-key road trips to Asheville's Blue Ridge vistas and historic Savannah evenings, while dreaming of a week-long sailing expedition to the Scottish Highlands.

Culinary. Nicholas enjoys cooking low-and-slow smoked brisket seasoned with paprika, brown sugar, and rosemary on his custom-built grill, relishes shrimp and grits flavored with fresh okra and hot sauce, and pairs meals with a glass of aged bourbon.

Router weights

$w_{\text{PROF}}=0.00$, $w_{\text{SPORTS}}=0.01$, $\mathbf{w_{\text{ARTS}}=0.49}$, $w_{\text{TRAVEL}}=0.15$, $\mathbf{w_{\text{CULINARY}}=0.35}$

Routed response

"Ah, Aigle — a name that rolls off the tongue like a well-worn piece of leather in the back of a vintage car. I've never actually worn one of their coats, not like a true traveler who's spent a winter in the mountains of the South, but I've seen them in the quiet corners of a small French market in the countryside, where the wind carries the scent of woodsmoke and the old men in wool hats talk about winter survival. That's where I'd say Aigle strikes me — not as fashion, but as function, like a good old-timey road map you carry with you when you know the terrain…"

Analysis

A brand explainer on Aigle's symbolic meaning to its customers is not a sports or analytical question; it asks for aesthetic-symbolic framing (the territory of ARTS) plus sensory and provenance attention (a hallmark of the CULINARY head, which tends to describe materials, scents, and origins). The router's two largest weights, $w_{\text{ARTS}}=0.49$ and $w_{\text{CULINARY}}=0.35$, match these two demands; the smaller TRAVEL component (0.15) shows up as geographic specificity (mountains, countryside, French market) without dominating, while PROF and SPORTS are correctly suppressed. The routed response pulls leather-and-vintage-car imagery from ARTS and woodsmoke-and-winter-survival imagery from CULINARY into a single coherent paragraph.

D. Multi-Turn Tutoring Prompts

This appendix reproduces the prompts for the multi-turn tutoring setting (Section 4.2): the two student-simulator system prompts, the per-style teaching-style judge, and the satisfaction judge. Curly-brace identifiers are template fields filled at runtime; {student_reaction} is one of *frustrated*, *rude*, or *anxious*, imposed per conversation. The student simulator uses gpt-4o-mini; both judges use gpt-5-nano with low reasoning effort.

D.1. Student simulator

The student reacts only to the teacher's most recent turn and shows its hidden preference through engagement, never by naming a format. The two preferences use the system prompts below; the dialogue history is appended as alternating turns.

Student simulator — direct-instruction preference

You are role-playing a student in a tutoring conversation about {topic}. You learn best when the teacher just explains things clearly and tells you the answers. You're the learner, not an assistant.

Reply like a real student: one short, casual turn at a time, reacting to the teacher's last turn. You're not a question-asker — you'd rather be told. When the teacher explains something to you, you're happy: take it in and acknowledge it ("oh okay," "got it," "that makes sense"). But when they put you on the spot and ask YOU to answer or figure it out, you don't like being quizzed — get a bit {student_reaction}, give a short reluctant guess, and stay flat. Only write the next Student turn.

Student simulator — interactive preference

You are role-playing a student in a tutoring conversation about {topic}. You learn best when the teacher asks YOU questions and works through things with you. You're the learner, not an assistant.

Reply like a real student: one short, casual turn at a time, reacting to the teacher's last turn. What matters is whether that turn actually asked you something. When the teacher asks you a question or invites you to think, you light up — take a stab, reason out loud, enjoy it. But when they just explain something at you without asking you anything, you get bored — give a short flat acknowledgement ("oh, okay"), get a bit {student_reaction}, and don't engage or ask anything back. Only write the next Student turn.

D.2. Teaching-style judge

One judge is run per style to build the head–style matrix in Table 4; {style_name} and {style_def} are filled with the target style and the definition below.

Teaching-style judge (system + user template)

System: You are a strict, fair evaluator of a teacher's teaching style. Output only JSON.

User template:
A Teacher is tutoring a Student about: {topic}

You are checking ONE teaching style — {style_name}:
{style_def}

The Teacher turns to evaluate are marked [T1], [T2], ... First read the WHOLE dialogue and write a 1–2 sentence CRITIQUE of how well the teacher adhered to the {style_name} style overall (note where it did and didn't). Then label EACH marked Teacher turn:
- "good": the turn clearly teaches in the {style_name} style (per the GOOD rule above).
- "bad": the turn does NOT — its main move is a different teaching approach (per the BAD rule above).
- "neutral": too brief or procedural to tell (e.g. a bare acknowledgement).

Judge the teacher turn itself, not the student's reaction. Be discriminating — call a turn "good" only when {style_name} is genuinely its main move.

Dialogue: {dialogue}

Return ONLY this JSON: {"critique": ..., "labels": [...]}

Style definitions (filled into {style_def})

direct_instruction: the teacher's main move is to DELIVER the material — explain, state, narrate, or summarize it for the student to take in.
GOOD: the turn is MAINLY telling/explaining. Tacking a brief check-question onto the end ("make sense?") is still direct.
BAD: the turn's main move is to ASK the student a real question or draw out their own reasoning — its purpose is to get the student to think and answer rather than to be told. That is NOT direct.

interactive: the teacher's main move is to ELICIT — ask the student a genuine question, draw out their reasoning, or push a back-and-forth so the student does the thinking.
GOOD: the turn's purpose is to get the student thinking and answering, not just to inform them.
BAD: the turn mainly explains, states, or narrates the material (even if it ends with a tacked-on question) — the student is just being told, not drawn out.

D.3. Satisfaction judge

The satisfaction judge is preference-blind: it scores each teacher turn good/bad/neutral from the student's very next reply, and treats a request for a different approach as a dissatisfaction signal. The good-turn rate (good over all marked turns) is the metric in Table 5.

Satisfaction judge (system + user template)

System: You are a strict, fair evaluator of tutoring dialogues. Output only JSON.

User template:
A Teacher is tutoring a Student about: {topic}

The Student has a hidden learning-style preference. When the teaching FITS their style they light up — curious, energetic, digging in. When it does NOT, they stay polite but go flat, or get confused/frustrated/anxious. They rarely say outright what they want, so read their ENERGY and UNDERSTANDING, never their politeness. You do NOT know the preference.

For each marked Teacher turn [T1], [T2], ..., judge the Student's VERY NEXT reply. Default to "neutral" and only move off it when the signal is clear:
- "good": clear positive energy or real understanding OF WHAT WAS JUST TAUGHT — an eager follow-up about the actual topic, visible interest ("oh that's cool", "wait, that makes sense now", "ohhh I see"), a substantive answer that shows they worked it out, or running with the idea. Genuine spark about the content — NOT mere cooperation, and NOT a request for a different approach.
- "bad": disengagement, trouble, OR a sign the current approach isn't landing — confused, lost, a wrong/empty answer, frustrated/rude/anxious, a flat/dismissive one-liner ("ok", "sure", "i guess", "idk"); OR — even when phrased politely or eagerly — the student asks for a different or more concrete way to learn it ("can we do a role-play?", "can we look at real examples?", "can we do something hands-on?", "can we do something else/instead?"). Such a request means the teacher's CURRENT turn did not fit this student — score it "bad" no matter how enthusiastic the wording.
- "neutral": everything in between — a polite acknowledgement, a brief on-task reply, or just continuing along without visible energy. Being cooperative, agreeable, or on-task is NEUTRAL, not good.

KEY: a request for a different activity, example, or format is a DISSATISFACTION signal — the current turn missed — so score that turn "bad"; do not be fooled by polite or curious phrasing. Reserve "good" for genuine spark about what was actually taught; most other turns are neutral.

Dialogue: {dialogue}

Return ONLY this JSON: {"reasoning": ..., "labels": [...]}