HSIR: HIDDEN-STATE INTRINSIC REWARDING FOR MULTILINGUAL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) exhibit strong yet uneven multilingual capabilities, motivating alignment methods that improve cross-lingual consistency. Existing approaches often derive rewards from translation models, multilingual encoders, or stronger language models, which may introduce a mismatch between the external reward space and the policy model's internal representations. We propose **HSIR**, a **H**idden-**S**tate **I**ntrinsic **R**ewarding framework for multilingual alignment. HSIR prepares reference answers offline before training and constructs an intrinsic reward during training by measuring the token-level similarity between the hidden states of generated responses and those of the corresponding reference answers. This reward is incorporated into Group Relative Policy Optimization (GRPO), eliminating the need for a separately trained reward model or a dedicated online evaluator. Experiments on multilingual open-ended generation show that HSIR consistently outperforms supervised fine-tuning and alternative semantic, cross-lingual, and translation-based reward methods. On Qwen2.5-3B-Instruct and Llama-3.2-3B-Instruct, HSIR achieves All-avg win rates of 49.02% and 66.32%, respectively, corresponding to improvements of 0.50 and 4.46 percentage points over the best-performing baselines. These results demonstrate that hidden representations provide an effective intrinsic reward for multilingual alignment. Code and data are available at `https://anonymous.4open.science/r/HSIR-4867`.

1 INTRODUCTION

Large language models (LLMs) can follow instructions, solve problems, and generate text in multiple languages (Chen et al., 2024; Xu et al., 2025). However, their capabilities remain uneven across languages, with substantially stronger performance in high-resource languages such as English than in many low-resource languages. This disparity has motivated multilingual post-training methods that seek to transfer useful capabilities from high-resource languages to low-resource languages (Shaham et al., 2024; Ji & Chen, 2025). Multilingual supervised fine-tuning (SFT) can improve performance in individual languages, but it does not directly enforce semantic consistency between responses across languages (Zhao et al., 2025). The central challenge is therefore to construct an optimization signal that rewards high-quality responses while preserving the intended semantics across languages.

Preference optimization provides a natural framework for this purpose because it can directly compare alternative responses according to a desired alignment criterion. MAPO (She et al., 2024) uses translation likelihood to measure consistency between responses in different languages. CM-Align (Zhang et al., 2025) constructs preference signals from cross-lingual semantic consistency and carefully selected reference responses. Other approaches, such as LIDR (Yang et al., 2025b), construct language-independent supervision from translated responses. Despite their differences, these methods generally rely on an auxiliary translation model, multilingual encoder, or stronger language model to evaluate generated responses. These evaluators produce rewards in representation spaces external to the policy model, which may not be well aligned with the representations used for generation. This raises the question of whether a dedicated external evaluator is necessary when the policy model already contains representations related to the desired response semantics.

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.

We therefore ask: can the hidden representations of the policy model itself provide a useful reference-based reward for multilingual alignment? To investigate this question, we propose HSIR, a **H**idden-**S**tate **I**ntrinsic **R**ewarding method for multilingual alignment. Before policy optimization, HSIR prepares reference answers offline, which remain fixed throughout training. For each prompt, it uses a frozen copy of the initial policy model to encode both a sampled response and its reference answer. HSIR then constructs a sequence-level reward by mean-pooling the hidden states of continuation tokens at each selected Transformer layer and computing the cosine similarity between the resulting response and reference representations. The layer-wise similarities are averaged to obtain a scalar reward, which is normalized within each response group and used by Group Relative Policy Optimization (GRPO) (Shao et al., 2024). Thus, HSIR does not require a separately trained reward head or a dedicated online translation model, multilingual encoder, or other external evaluator, while still using the provided reference answers as supervision.

We evaluate HSIR on Qwen2.5-3B-Instruct and Llama-3.2-3B-Instruct using multilingual open-ended generation and additionally examine its performance on multilingual mathematical reasoning. The alignment data cover five languages, while the evaluation also includes languages that are not used during alignment training. On OMGEval (Liu et al., 2024), HSIR achieves All-avg win rates of 49.02% on Qwen2.5-3B-Instruct and 66.32% on Llama-3.2-3B-Instruct. Relative to the strongest reported baselines, these results represent gains of 0.50 and 4.46 percentage points, respectively, indicating a substantially larger benefit on the Llama backbone. On multilingual mathematical reasoning, HSIR achieves the best average accuracy among the evaluated methods for Qwen, whereas its result on Llama remains below several reward-based baselines. Together, these results show that hidden-state similarity can serve as a competitive reference-based reward for multilingual instruction following, while also indicating that its transfer benefit depends on the model and task. We further analyze which Transformer layers contribute to the reward, how layer-wise similarities should be aggregated, and how the amount of alignment data affects performance.

2 Related Work

2.1 External Reward Modeling and Preference Optimization

Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) aligns large language models (LLMs) with human preferences by optimizing against learned reward models, while Direct Preference Optimization (DPO) (Rafailov et al., 2023) and subsequent preference optimization methods (Azar et al., 2024; Tang et al., 2024; Meng et al., 2024; Hong et al., 2024) directly optimize from preference data. LLM-based judges further provide flexible reward signals for helpfulness, correctness, relevance, and multilingual evaluation (Liu et al., 2023; Zheng et al., 2023; Hada et al., 2024; Liu et al., 2024), but these approaches generally rely on human annotations, auxiliary reward models, or additional LLM inference. In multilingual alignment, external models are likewise used to construct cross-lingual rewards: multilingual instruction tuning transfers knowledge from resource-rich languages (Muennighoff et al., 2023), while MAPO (She et al., 2024), LIDR (Yang et al., 2025b), and Implicit Cross-Lingual Rewarding (Yang et al., 2025a) leverage translation models, translated responses, or aligned models to derive cross-lingual preference signals. Despite their effectiveness, these approaches define rewards outside the policy, which may create a mismatch with its internal representation.

2.2 Internal Representations for Reward Modeling

Recent work has increasingly investigated the information encoded in the internal representations of language models. Hidden states have been shown to contain rich information about linguistic structure, factual knowledge, and latent model behavior (Ferrando et al., 2024; Li et al., 2023a; Burns et al., 2023). These representations have been exploited for applications including representation engineering (Zou et al., 2023), knowledge editing (Meng et al., 2022), and knowledge transfer or distillation (Dasgupta & Cohn, 2025; Wu et al., 2024). Collectively, these studies suggest that intermediate representations encode information that is highly relevant to the behavior and capabilities of the underlying language model, motivating their use beyond conventional next-token prediction. Building on this observation, recent work has explored internal representations as direct signals for evaluating generated responses. SWIFT (Guo et al., 2026) uses hidden states and logits from the task-performing LLM to construct lightweight intrinsic rewards for Best-of-N sampling,

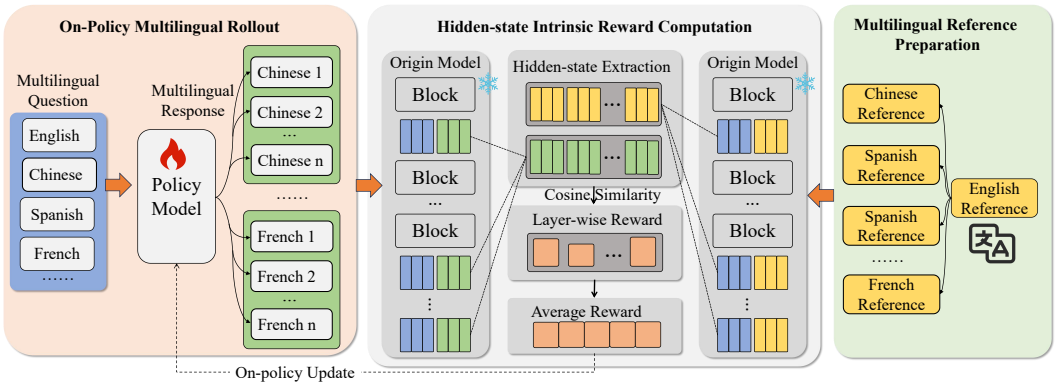


Figure 1: Overview of the proposed HSIR framework. The pipeline consists of three stages: (1) **Multilingual Reference Preparation**, which translates English references into target languages; (2) **On-Policy Multilingual Rollout**, where the policy generates response groups; and (3) **Hidden-State Intrinsic Reward Computation**, which compares sampled responses with target-language references in hidden-state space to obtain intrinsic rewards for GRPO optimization.

reducing the cost of conventional reward models. Our work extends this direction to multilingual alignment: rather than learning a separate reward model, HSIR directly measures hidden-state similarity between generated responses and their references within the policy model, and incorporates this intrinsic reward into GRPO for on-policy optimization.

3 Methodology

HSIR uses hidden-state similarity as an intrinsic reward for cross-lingual alignment. Let $\mathcal{D} = \{(x_i^{(l)}, y_i^{\star(l)})\}_{i=1}^{|\mathcal{D}|}$ denote the training set, where $x_i^{(l)}$ is a prompt in language l and $y_i^{\star(l)}$ is its target-language reference, obtained by translating an English reference $\bar{y}_i$ during offline data preparation. The reference provides a semantic and reasoning target without requiring the sampled response to match its wording token by token. For each prompt, sampled continuations are compared with this target in the language model's own representation space. Their hidden-state similarities are converted into group-relative rewards and optimized with standard GRPO. The reward calculation uses neither an online translation model nor a separately trained reward model.

3.1 Hidden-State Representation Extraction

For each training instance $(x_i^{(l)}, y_i^{\star(l)})$, the rollout policy samples a group of responses:

$$y_{i,g}^{(l)} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x_i^{(l)}), \quad g = 1, \ldots, G. \tag{1}$$

We apply the tokenizer's chat template to $x_i^{(l)}$ and append either a sampled response or $y_i^{\star(l)}$. Both sequences therefore share the same prompt formatting, while their continuation tokens are encoded independently by a frozen, inference-mode scoring model with parameters $\bar{\theta}$; $\bar{\theta}$ remains fixed throughout optimization. Let $y \in \{y_{i,g}^{(l)}, y_i^{\star(l)}\}$ have T_y continuation tokens, and let $H_{\bar{\theta}}^{(k)}(x_i^{(l)}, y) \in \mathbb{R}^{T_y \times d}$ denote the corresponding continuation states at layer k. We mean-pool them as

$$h_{\bar{\theta}}^{(k)}(x_i^{(l)}, y) = \frac{1}{T_y} \sum_{t=1}^{T_y} H_{\bar{\theta},t}^{(k)}(x_i^{(l)}, y). \tag{2}$$

Only continuation tokens are pooled, excluding the shared prompt states that would otherwise dominate the comparison and obscure differences between responses. Mean pooling also produces a fixed-dimensional representation for responses with different lengths. An empty continuation is replaced by an EOS or UNK token so that $T_y \geq 1$.

3.2 LAYER-WISE HIDDEN-STATE INTRINSIC REWARD

Let $\mathcal{K}$ denote the selected Transformer layers. For each layer $k \in \mathcal{K}$, we compute the cosine similarity between the sampled response and its corresponding target-language reference response:

$$r_{i,g}^{(k,l)} = \frac{\left(h_{\bar{\theta}}^{(k)}(x_i^{(l)}, y_{i,g}^{(l)})\right)^\top h_{\bar{\theta}}^{(k)}(x_i^{(l)}, y_i^{\star(l)})}{\left\|h_{\bar{\theta}}^{(k)}(x_i^{(l)}, y_{i,g}^{(l)})\right\|_2 \left\|h_{\bar{\theta}}^{(k)}(x_i^{(l)}, y_i^{\star(l)})\right\|_2}. \tag{3}$$

The final HSIR reward is obtained by uniformly averaging the layer-wise similarities:

$$R_{i,g}^{(l)} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} r_{i,g}^{(k,l)}. \tag{4}$$

We use uniform layer averaging; when no subset is specified, $\mathcal{K}$ contains all Transformer layers. This aggregates information distributed across the model's hierarchy, from relatively local linguistic features to higher-level semantic and reasoning features, without introducing additional layer-specific parameters. Because each term is cosine-normalized, the reward is insensitive to vector norms and requires only forward inference in the model's latent space.

3.3 GRPO-BASED ON-POLICY OPTIMIZATION

For each prompt in language l, the HSIR rewards are standardized within the sampled response group. This makes the update depend on the relative quality of responses sampled for the same prompt rather than on the absolute scale of cosine similarities, which can vary across prompts and languages:

$$\widehat{A}_{i,g}^{(l)} = \frac{R_{i,g}^{(l)} - \text{mean}\left(\{R_{i,j}^{(l)}\}_{j=1}^{G}\right)}{\text{std}\left(\{R_{i,j}^{(l)}\}_{j=1}^{G}\right) + \delta}. \tag{5}$$

Responses with above-average similarity receive positive advantages, whereas below-average responses receive negative advantages. The same response-level advantage is assigned to every token in a response, so the policy reinforces or suppresses the complete sampled continuation. For each response token, define the importance ratio and its clipped version as:

$$\rho_{i,g,t}^{(l)}(\theta) = \frac{\pi_\theta(y_{i,g,t}^{(l)} \mid x_i^{(l)}, y_{i,g,<t}^{(l)})}{\pi_{\theta_{\text{old}}}(y_{i,g,t}^{(l)} \mid x_i^{(l)}, y_{i,g,<t}^{(l)})}, \tag{6}$$

$$\bar{\rho}_{i,g,t}^{(l)}(\theta) = \text{clip}\left(\rho_{i,g,t}^{(l)}(\theta), 1-\epsilon, 1+\epsilon\right). \tag{7}$$

The policy is optimized using the standard clipped GRPO objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{g=1}^{G}\frac{1}{T_{i,g}}\sum_{t=1}^{T_{i,g}}\min\left(\rho_{i,g,t}^{(l)}(\theta)\widehat{A}_{i,g}^{(l)}, \bar{\rho}_{i,g,t}^{(l)}(\theta)\widehat{A}_{i,g}^{(l)}\right)\right] - \beta D_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}}\right). \tag{8}$$

Here, π_{ref} is the fixed reference policy and β controls the KL penalty, which limits unnecessary deviation from the pretrained behavior. The rollout policy $\pi_{\theta_{\text{old}}}$ is updated on-policy, whereas the scoring model with parameters $\bar{\theta}$ remains frozen. The GRPO update is otherwise unchanged; HSIR contributes the construction of $R_{i,g}^{(l)}$ from hidden-state similarity to $y_i^{\star(l)}$.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Base Models. We use two widely adopted multilingual instruction-following LLMs, Qwen2.5-3B-Instruct Qwen et al. (2025) and Llama-3.2-3B-Instruct Grattafiori et al. (2024). Despite broad multilingual pretraining, both exhibit uneven performance across the evaluated languages. Our goal is to improve multilingual performance through alignment.

Datasets. Following Zhang et al. (2025), we construct a 1.5K multilingual instruction-following training set from AlpacaEval Li et al. (2023b), covering five languages: English (en), Chinese (zh), Spanish (es), French (fr), and Russian (ru). For evaluation, we use OMGEval Liu et al. (2024) to assess general instruction following. We further evaluate multilingual mathematical reasoning on MGSM8K Shi et al. (2023), which covers ten languages: en, zh, es, fr, ru, Bengali (bn), German (de), Thai (th), Swahili (sw), and Japanese (ja).

Evaluation. For general instruction following, we evaluate against GPT-3.5-turbo on OMGEval using win rate (WR). We extend the AlpacaEval 2.0 toolkit to support multilingual evaluation. For mathematical reasoning, we report answer accuracy (Accuracy, %) on MGSM8K, where a prediction is correct if its final answer matches the ground truth.

Implementation Details. Training is conducted on four GPUs with bfloat16 mixed precision and CPU offloading for memory efficiency. We use vLLM for generation, sampling four responses per prompt with a temperature of 1.0. Both input and output lengths are capped at 2048 tokens. We use a batch size of 2, four gradient accumulation steps, a learning rate of 1×10^{-6}, and a warmup ratio of 0.01. Models are trained for one epoch per iteration. DeepSpeed ZeRO Stage 3 is used for further memory optimization, with $\beta = 0.001$.

4.2 Baselines

MAPO and MAPO$_{\text{cross}}$. MAPO She et al. (2024) uses translation likelihood to construct a cross-lingual reward. We use nllb-200-distilled-600M NLLB Team (2024) to measure the likelihood of the English reference conditioned on a generated non-dominant-language response. MAPO follows the original reward and optimization formulation, while MAPO$_{\text{cross}}$ uses the English reference as the target and optimizes the translation reward with GRPO for a controlled comparison.

LIDR. LIDR Yang et al. (2025b) constructs preference data using translation-based supervision. For non-dominant languages, the translated English response is used as the chosen response and the original response as the rejected response; the roles are reversed for English.

SFT$_{\text{multi}}$. SFT$_{\text{multi}}$ Shaham et al. (2024) performs supervised fine-tuning on multilingual instruction-response pairs, providing a non-reward-based baseline for multilingual alignment.

LLM$_{\text{judge}}$. LLM$_{\text{judge}}$ Wu et al. (2025) uses an LLM-based judge as the reward for multilingual alignment, following the LLM-as-a-Judge formulation. We use Qwen2.5-7B-Instruct Qwen et al. (2025) as the judge model to score the generated responses, and optimize the judge reward with GRPO to align the optimization framework with our method.

CM-Align, CM-Align$_{\text{cross}}$, and CM-Align$_{\text{multi}}$. CM-Align Zhang et al. (2025) uses semantic similarity as a reward for multilingual alignment. We use gte-multilingual-base Zhang et al. (2024) to compute the similarity between generated and reference responses. CM-Align follows the original reward formulation, while CM-Align$_{\text{cross}}$ compares non-dominant-language responses with English references and CM-Align$_{\text{multi}}$ compares them with same-language references. For the two variants, the similarity scores are optimized with GRPO to align the optimization framework with our method.

4.3 Main Results

Table 1 reports multilingual instruction-following results on OMGEval, where HSIR achieves the best overall performance for both base models. On Qwen2.5-3B-Instruct, HSIR reaches an All-avg win rate of 49.02%, improving over the original model by 21.58 points and surpassing both SFT$_{\text{multi}}$ (28.52%) and the strongest GRPO-based baselines MAPO$_{\text{cross}}$ (47.68%), CM-Align$_{\text{cross}}$ (48.52%), and LLM$_{\text{judge}}$ (46.57%). The advantage is larger on Llama-3.2-3B-Instruct, where HSIR obtains 66.32% All-avg, exceeding the original model by 20.31 points and the best baseline (CM-Align$_{\text{cross}}$, 61.86%) by 4.46 points. These results show that hidden-state similarity provides an effective intrinsic reward for multilingual alignment, outperforming both translation-based and externally defined semantic-similarity rewards under the same optimization framework.

Language-level results confirm that HSIR improves multilingual performance broadly rather than concentrating gains on a single language. On Qwen2.5-3B-Instruct, it achieves the best scores on Chinese (85.20%), French (50.06%), and Russian (47.76%), yielding the strongest in-domain average (61.45%) while remaining competitive on English and Spanish. On Llama-3.2-3B-Instruct, it

Method	In-Domain						Out-of-Domain			All-avg
	en	zh	es	fr	ru	**avg**	bn	de	**avg**	
Qwen2.5-3B	48.14	19.03	14.18	17.35	24.88	24.72	29.94	30.37	30.16	27.44
+ $\text{SFT}_{\text{multi}}$	63.98	57.96	22.57	24.44	28.79	39.55	8.71	26.27	17.49	28.52
Methods without labels										
+ MAPO	59.81	59.64	31.72	30.53	37.81	43.90	31.06	50.68	40.87	42.38
+ LIDR	68.76	76.37	16.98	21.08	21.58	40.95	12.17	30.93	21.55	31.25
+ CM-Align	62.67	64.61	31.84	33.02	40.05	46.44	33.54	52.30	42.92	44.68
Methods with labels										
+ $\text{LLM}_{\text{judge}}$	78.88	83.83	**47.14**	43.16	46.52	59.91	14.78	51.68	33.23	46.57
+ $\text{MAPO}_{\text{cross}}$	**79.07**	85.07	45.65	44.71	49.00	60.70	18.14	51.18	34.66	47.68
+ $\text{CM-Align}_{\text{cross}}$	78.07	83.58	45.77	46.14	46.89	60.09	22.36	51.55	36.96	48.52
+ $\text{CM-Align}_{\text{multi}}$	77.14	82.96	44.15	42.16	45.09	58.30	**22.61**	**52.80**	**37.71**	48.01
+ HSIR (ours)	78.70	**85.20**	45.52	**50.06**	**47.76**	**61.45**	21.49	51.68	36.59	**49.02**
Llama-3.2-3B	75.40	17.29	33.15	30.97	14.76	34.31	52.17	63.23	57.70	46.01
+ $\text{SFT}_{\text{multi}}$	78.20	20.58	37.86	32.09	12.19	36.18	27.70	54.61	41.16	38.67
Methods without labels										
+ MAPO	77.70	33.21	42.79	39.05	21.77	42.90	60.25	71.18	65.72	54.31
+ LIDR	83.35	36.32	46.89	42.41	21.89	46.17	49.57	70.56	60.07	53.12
+ CM-Align	76.34	31.97	46.89	39.55	23.01	43.55	69.81	73.60	71.71	57.63
Methods with labels										
+ $\text{LLM}_{\text{judge}}$	87.45	44.90	58.41	53.62	22.42	53.36	59.13	77.27	68.20	60.78
+ $\text{MAPO}_{\text{cross}}$	87.08	47.14	58.21	54.85	24.88	54.43	63.98	73.79	68.89	61.66
+ $\text{CM-Align}_{\text{cross}}$	86.21	44.90	57.34	51.49	22.39	52.47	65.22	77.27	71.25	61.86
+ $\text{CM-Align}_{\text{multi}}$	84.72	42.41	54.35	48.63	**26.49**	51.32	58.88	74.29	66.59	58.95
+ HSIR (ours)	**89.69**	**47.64**	**63.18**	**57.46**	25.12	**56.62**	**73.79**	**78.26**	**76.03**	**66.32**

Table 1: Win rate (WR) results (%) on the OMGEval benchmark.

leads on English (89.69%), Chinese (47.64%), Spanish (63.18%), and French (57.46%) and attains the best in-domain (56.62%) and OOD (76.03%) averages, indicating strong transfer to the unseen languages Bengali and German. On Qwen, HSIR's OOD average is only slightly below that of $\text{CM-Align}_{\text{multi}}$, but its in-domain and overall gains are substantially larger. These results suggest that HSIR captures semantic signals that transfer across both training and unseen languages.

5 Analysis

5.1 Reward Dynamics

Figure 2 compares reward dynamics over the first 125 optimization steps. Since the rewards operate on different scales, each trajectory is min–max normalized, so the comparison focuses on temporal behavior rather than magnitude. HSIR shows a sustained upward trend, rising from 0.33 (first 25 steps) to 0.74 (last 25 steps) with the strongest linear trend (slope 0.0043, $r = 0.74$), indicating the signal remains informative as the policy evolves rather than saturating. $\text{CM-Align}_{\text{cross}}$ improves comparably ($0.35 \rightarrow 0.74$), while $\text{CM-Align}_{\text{multi}}$ rises more weakly ($0.38 \rightarrow 0.58$). In contrast, $\text{MAPO}_{\text{cross}}$ oscillates with little net change ($0.25 \rightarrow 0.24$, $r \approx 0$), and $\text{LLM}_{\text{judge}}$ improves steadily ($0.28 \rightarrow 0.61$, slope 0.0035) but plateaus below HSIR, showing that even a strong external judge yields weaker dynamics than the intrinsic reward. Overall, HSIR combines a strong late-stage reward with a stable trajectory, providing a more consistent learning direction than translation-likelihood rewards, consistent with the win rates in Table 1.

5.2 Layer Analysis

We investigate the effect of Transformer depth on the intrinsic reward by comparing three contiguous eight-layer subsets with the all-layer HSIR variant. Figure 3 reports WR on OMGEval for Llama-3.2-3B-Instruct. All variants outperform the original model (46.01% All-avg), confirming that hidden-state similarity remains effective with restricted layer ranges. The all-layer variant achieves the best overall performance (66.32%), improving by 20.31 points over the original model and 0.86 points over the bottom-layer variant, while also obtaining the highest in-domain average

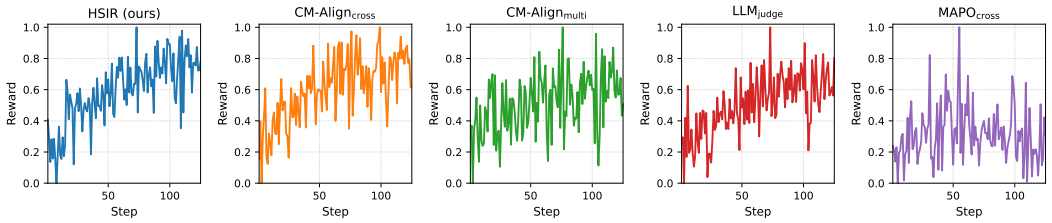


Figure 2: Normalized reward trajectories over the training steps.

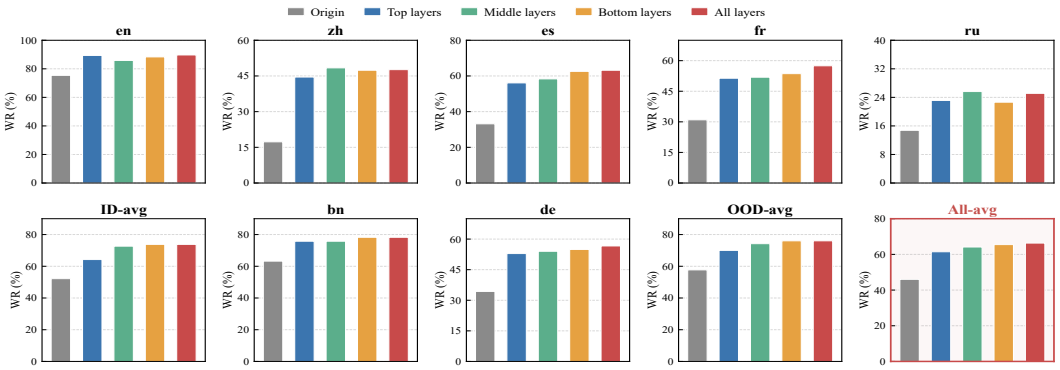


Figure 3: Layer-selection ablation for HSIR on Llama-3.2-3B-Instruct.

(56.62%). The bottom-layer variant remains competitive (65.46%) and matches the all-layer variant on OOD average (76.03%). Performance is not monotonic with depth: middle layers perform best on Chinese and Russian, bottom layers lead on Bengali and German, while top layers favor English but yield the lowest OOD average among adapted variants. Overall, these results suggest that lower and middle layers capture useful cross-lingual transfer signals, while upper layers contribute more to dominant-language behavior, making all-layer aggregation a robust choice across languages.

5.3 Reward Aggregation

HSIR aggregates layer-wise similarities by computing cosine similarity at each selected layer and averaging the scores (Eq. 3–4). We compare this strategy with "Layer-Concat", which mean-pools and concatenates hidden states across layers before computing a single cosine similarity. As shown in Figure 4, "Layer-Averaged (HSIR)" consistently outperforms "Layer-Concat", improving the All-avg win rate from 62.39% to 66.32%, with gains broadly distributed across languages and consistent across in-domain and out-of-domain averages. This suggests that layer-wise averaging provides a more robust reward signal by preserving independent alignment evidence across heterogeneous layers, whereas concatenation couples representations and increases sensitivity to scale and geometry differences.

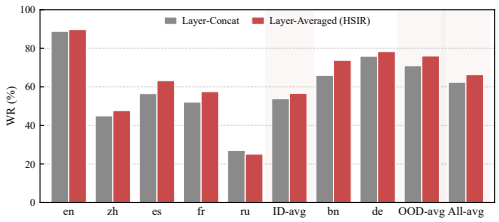


Figure 4: Comparison of reward aggregation strategies for HSIR.

5.4 Generalization to Unseen Tasks

Table 2 evaluates transfer to multilingual mathematical reasoning on MGSM8K. For Qwen2.5-3B-Instruct, HSIR achieves the highest average accuracy (56.08%), improving over the original model by 9.20 points and surpassing all baselines, including MAPO$_{cross}$ (54.56%) and CM-Align$_{multi}$ (55.80%); it also leads on five of the ten languages, showing that the hidden-state reward transfers beyond instruction following to mathematical reasoning. For Llama-3.2-3B-Instruct, HSIR

Method	In-Domain Language					Out-of-Domain Language					avg
	en	zh	es	fr	ru	bn	de	th	sw	ja	
Qwen2.5-3B	79.20	56.00	58.40	56.80	62.80	22.40	44.40	48.40	2.80	37.60	46.88
+ $\text{SFT}_{\text{multi}}$	78.00	69.60	57.60	51.60	57.20	27.20	41.60	44.00	5.60	34.00	46.64
+ $\text{LLM}_{\text{judge}}$	78.40	**72.00**	64.00	58.00	63.20	32.00	56.80	56.00	4.00	51.60	53.60
+ $\text{MAPO}_{\text{cross}}$	77.20	71.20	61.60	62.80	65.60	35.20	56.40	56.80	4.80	54.00	54.56
+ $\text{CM-Align}_{\text{cross}}$	82.00	71.60	62.40	60.80	65.60	35.60	56.00	56.80	**6.80**	52.00	54.96
+ $\text{CM-Align}_{\text{multi}}$	78.80	70.80	**65.60**	62.00	65.60	34.00	**61.60**	58.40	6.40	54.80	55.80
+ HSIR (ours)	**82.80**	69.60	64.00	**63.60**	**66.80**	**35.60**	55.20	**62.00**	5.60	**55.60**	**56.08**
Llama-3.2-3B	71.60	59.60	**67.20**	**60.80**	62.40	44.00	**63.20**	58.80	51.20	49.20	58.80
+ $\text{SFT}_{\text{multi}}$	70.00	56.80	53.20	57.20	54.40	49.20	57.60	55.20	46.80	48.40	54.88
+ $\text{LLM}_{\text{judge}}$	64.40	64.40	59.60	53.60	64.80	56.40	58.40	**66.80**	54.80	**59.20**	**60.24**
+ $\text{MAPO}_{\text{cross}}$	70.00	65.20	52.80	52.40	61.20	**58.00**	58.80	63.60	52.40	55.60	59.00
+ $\text{CM-Align}_{\text{cross}}$	70.80	**67.60**	54.00	51.20	**66.40**	56.80	56.80	65.20	56.00	52.80	59.76
+ $\text{CM-Align}_{\text{multi}}$	**72.00**	66.80	57.20	53.60	65.60	54.00	61.20	59.60	54.00	56.40	60.04
+ HSIR (ours)	70.00	64.80	58.80	57.60	65.60	53.20	58.00	57.20	**56.80**	54.80	59.68

Table 2: Accuracy (%) on MGSM8K across in-domain and out-of-domain languages.

attains 59.68%, slightly above the original model (58.80%) but below $\text{LLM}_{\text{judge}}$ (60.24%) and $\text{CM-Align}_{\text{multi}}$ (60.04%). This indicates the benefit is task- and model-dependent: it transfers effectively for Qwen but remains more limited on Llama.

5.5 Completion Analysis

Figure 5 shows that increasing the number of sampled completions improves HSIR's multilingual performance on Llama-3.2-3B-Instruct: All-avg rises from 46.01% for the original model to 59.53%, 66.32%, and 67.16% for `com_2`, `com_4`, and `com_8`, respectively. The largest gain occurs with `com_2`, followed by diminishing returns; although some languages peak at `com_4`, the out-of-domain average continues to improve slightly with `com_8`. The 0.84-point gain from `com_4` to `com_8` further indicates that most of the benefit is captured with four completions. These results support `com_4` as a practical trade-off between performance and sampling cost.

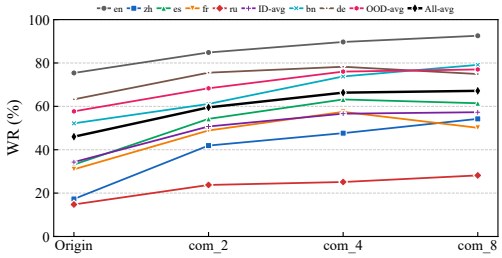


Figure 5: Effect of the number of completions.

5.6 Data Scale Analysis

To study how the amount of alignment data affects HSIR, we train Llama-3.2-3B-Instruct with different fractions of the alignment set while keeping the training procedure and evaluation protocol fixed. Table 3 shows that increasing the data scale consistently improves the overall win rate. Using only 20% of the data raises the All-avg score from 46.01% to 54.11% (+8.10 points), while using the full dataset reaches 66.32% (+20.31 points). The gains are evident for both in-domain and out-of-domain languages: the ID average increases from 34.31% to 56.62%, and the OOD average increases from 57.70% to 76.03%. This improvement on Bengali and German, which are not included in alignment training, indicates that additional multilingual alignment examples strengthen transfer beyond the training languages. Although the marginal gains fluctuate across data scales and languages (e.g., Russian temporarily decreases at 80%), the aggregate trend is strongly positive, supporting the benefit of using more alignment data for HSIR.

5.7 Logit-Based Training

We further compare HSIR's hidden-state similarity reward with a logit-distribution reward on Llama-3.2-3B-Instruct. The logit reward averages next-token probability distributions over each continuation and computes one minus their normalized Jensen–Shannon Lin (1991) divergence. As

	In-Domain						Out-of-Domain			
	en	zh	es	fr	ru	**ID-avg**	bn	de	**OOD-avg**	**All-avg**
Origin	75.40	17.29	33.15	30.97	14.76	34.31	52.17	63.23	57.70	46.01
20%	80.87	33.33	47.64	41.29	22.01	45.03	56.65	69.75	63.20	54.11
40%	83.11	39.80	50.00	46.33	23.51	48.55	57.14	71.93	64.54	56.54
60%	84.97	41.79	57.78	50.25	24.50	51.86	67.83	75.40	71.62	61.74
80%	86.83	46.52	61.94	53.98	22.26	54.31	69.32	76.89	73.11	63.71
100%	**89.69**	**47.64**	**63.18**	**57.46**	**25.12**	**56.62**	**73.79**	**78.26**	**76.03**	**66.32**

Table 3: Effect of data scale on OMGEval.

	In-Domain						Out-of-Domain			
	en	zh	es	fr	ru	**ID-avg**	bn	de	**OOD-avg**	**All-avg**
Origin	75.40	17.29	33.15	30.97	14.76	34.31	52.17	63.23	57.70	46.01
logit	81.45	49.13	58.53	53.98	23.51	53.32	66.04	74.91	70.48	61.90
hidden	**89.69**	**47.64**	**63.18**	**57.46**	**25.12**	**56.62**	**73.79**	**78.26**	**76.03**	**66.32**

Table 4: Comparison of hidden-state and logit-distribution rewards (WR, %).

shown in Table 4, both rewards substantially improve over the original model, confirming the effectiveness of intrinsic similarity signals for multilingual alignment. The hidden-state reward achieves a higher All-avg of 66.32%, compared with 61.90% for the logit reward, with improvements in both in-domain and out-of-domain languages. Notably, hidden-state similarity yields a larger gain on OOD languages (76.03% vs. 70.48%), suggesting that hidden representations provide a more effective signal for capturing cross-lingual correspondence and generalizing beyond the training languages.

5.8 Case Study

As shown in Table 5, both models correctly mention the actors listed in the reference, including Tom Hanks, Meryl Streep, Leonardo DiCaprio, and Brad Pitt. However, the original model additionally introduces a specific statement about DiCaprio beginning his career in television series, which is not contained in the reference. In contrast, the HSIR-optimized model generates a sentence that closely paraphrases the reference's final statement, preserving its meaning that the actors began their careers at different times and through different opportunities before gaining recognition in the film industry. This example suggests that HSIR improves semantic alignment with the reference while maintaining fluent target-language generation, using hidden-state similarity as an intrinsic reward without requiring a separately trained multilingual reward model.

Spanish Question: ¿Cuáles son los nombres de algunos actores famosos que comenzaron sus carreras?

Spanish Reference:
Hay muchos actores famosos que comenzaron sus carreras de diferentes maneras. Algunos ejemplos conocidos son Tom Hanks, Meryl Streep, Leonardo DiCaprio y Brad Pitt. Comenzaron sus carreras en diferentes momentos y mediante distintas oportunidades, pero cada uno llegó a ser un actor muy reconocido en la industria cinematográfica. Sus carreras abarcan una amplia variedad de películas y géneros, y han recibido un reconocimiento considerable por su trabajo.

Spanish Response:
[Origin] Sí, hay muchos actores famosos que comenzaron sus carreras de diferentes maneras. Algunos ejemplos conocidos son Tom Hanks, Meryl Streep, Leonardo DiCaprio y Brad Pitt. Leonardo DiCaprio, por ejemplo, comenzó su carrera principalmente en series de televisión antes de convertirse en una de las estrellas de cine más reconocidas de Hollywood.
[HSIR] Sí, hay muchos actores famosos que comenzaron sus carreras de diferentes maneras. Algunos ejemplos conocidos son Tom Hanks, Meryl Streep, Leonardo DiCaprio y Brad Pitt. Cada uno comenzó su trayectoria en diferentes momentos y mediante distintas oportunidades, pero posteriormente alcanzó un gran reconocimiento en la industria cinematográfica.

Table 5: A qualitative example from OMGEval comparing the original model with HSIR.

6 Conclusion

We presented HSIR, a hidden-state intrinsic reward framework for multilingual alignment. By measuring similarity between generated responses and offline-prepared target-language references in the policy model's hidden space, HSIR provides an intrinsic reward for GRPO. Experiments on OMGEval and MGSM8K show that HSIR improves multilingual alignment across both backbones, demonstrating the practical value of policy hidden states as an alignment signal.

AI USE STATEMENT

We used AI-based tools solely for language polishing, proofreading, and LaTeX formatting checks. AI tools were not used to develop the scientific ideas, methodology, experiments, analyses, or conclusions of this work. All AI-assisted suggestions were critically reviewed and verified by the authors, who remain fully responsible for the content and integrity of the manuscript.

REFERENCES

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International conference on artificial intelligence and statistics*, pp. 4447–4455. PMLR, 2024.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7001–7016, 2024.

Sayantan Dasgupta and Trevor Cohn. Improving language model distillation through hidden state matching. In *International Conference on Learning Representations*, volume 2025, pp. 19035–19049, 2025.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R Costa-Jussà. A primer on the inner workings of transformer-based language models. *arXiv preprint arXiv:2405.00208*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jizhou Guo, Zhaomin Wu, Hanchen Yang, and Philip S Yu. Mining intrinsic rewards from llm hidden states for efficient best-of-n sampling. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 371–382, 2026.

Rishav Hada, Varun Gumma, Adrian De Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. Are large language model-based evaluators the solution to scaling up multilingual evaluation? In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1051–1070, 2024.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189, 2024.

Shaoxiong Ji and Pinzhen Chen. How many languages make good multilingual instruction tuning? a case study on bloom. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 2575–2581, 2025.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Vi'egas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations*, 2023a.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models, 2023b.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Yang Liu, Meng Xu, Shuo Wang, Liner Yang, Haoyu Wang, Zhenghao Liu, Cunliang Kong, Yun Chen, Maosong Sun, and Erhong Yang. Omgeval: An open multilingual generative evaluation benchmark for large language models. *arXiv preprint arXiv:2402.13524*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 1: Long papers)*, pp. 15991–16111, 2023.

NLLB Team. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841—846, June 2024. ISSN 0028-0836. doi: 10.1038/s41586-024-07335-x.

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

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2304–2317, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. Mapo: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10015–10027, 2024.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language Models Are Multilingual Chain-of-Thought Reasoners. In *International Conference on Learning Representations*, 2023. URL https://mlanthology.org/iclr/2023/shi2023iclr-language/.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Rémi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Ávila Pires, and Bilal Piot. Generalized preference optimization: a unified approach to offline alignment. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason E Weston, and Sainbayar Sukhbaatar. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11548–11565, 2025.

Zhaomin Wu, Jizhou Guo, Junyi Hou, Bingsheng He, Lixin Fan, and Qiang Yang. Model-based differentially private knowledge transfer for large language models. *arXiv preprint arXiv:2410.10481*, 2024.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362, 2025.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. Implicit cross-lingual rewarding for efficient multilingual preference alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21125–21147, 2025a.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. Language imbalance driven rewarding for multilingual self-improving. In *International Conference on Learning Representations*, volume 2025, pp. 43911–43936, 2025b.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 conference on empirical methods in natural language processing: industry track*, pp. 1393–1412, 2024.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. Cm-align: Consistency-based multilingual alignment for large language models. In *EMNLP (Findings)*, pp. 25689–25702, 2025.

Weixiang Zhao, Yulin Hu, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, et al. Mpo: Multilingual safety alignment via reward gap optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23564–23587, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.