From Explicit to Implicit: A Theoretical Framework and Transfer Method for Preference Internalization in Language Models

Binrui Wang[1,2] **Yongping Du**[1,2]* **Yu Pei**[1,2] **Zikai Wang**[1,2]
[1]College of Computer Science, Beijing University of Technology, Beijing, China
[2]Beijing Artificial Intelligence Institute, Beijing University of Technology, Beijing, China
*ypdu@bjut.edu.cn

Abstract

Transforming explicit preference signals into implicit and parameterized behaviors is pivotal for enabling prompt-free, human-aligned generation and improving the usability, efficiency and robustness of large language models. However, existing methods align model preference well but still rely on explicit user instructions to convey specific preferences, leading to cumbersome user experiences and undermining natural, frictionless interaction with the model. To fill the gap between explicit and implicit preference representation, this paper introduces a theoretical framework that establishes both necessary and sufficient conditions for effective preference recognition. Based on this framework, we propose a novel Two-Stage Progressive Preference Transfer (TSPPT) method, which decomposes preference internalization into two manageable stages: preference representation learning and preference internalization transfer. The proposed method fills the gap between explicit and implicit preferences while maintaining the model's general capabilities. The experiments across multiple models (Qwen2.5, Qwen3, Llama-3.2, DeepSeek-R1-Distill) and datasets (UltraFeedback, HelpSteer) demonstrate superior performance. The proposed method achieves 79.2% win rate on UltraFeedback (vs. 59.2-67.6% for baselines), substantial improvements on MT-Bench (7.86 vs. 7.34 for best baseline), and significant reductions in implicit social bias (0.165 vs. 0.185-0.325 for baselines). Notably, the method maintains comparable performance between implicit and explicit settings, confirming successful preference internalization.[1]

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across diverse tasks (DeepSeek-AI et al., 2025; OpenAI et al., 2024), yet aligning their outputs with human preferences remains a fundamental challenge (WANG et al., 2023; Lee et al., 2024b; Ji et al., 2024b; Thakkar et al., 2024; Sun et al., 2023; Lee et al., 2024a). Human preferences are inherently implicit, as it is not common to explicitly state "please use a polite tone, provide concise answers, and avoid technical speech" when interacting with a human assistant. An aligned intelligence should make such preferences the default rather than requiring explicit instructions in each interaction. (Lerner et al., 2024). However, current models rely on prompts that include explicit preference instructions, which leads to uneven behavior when such instructions are omitted and raises a recurring burden on users to restate their preferences (Chen et al., 2025; Cao et al., 2024). These limitations pose a significant problem: How can explicit preference instructions in the prompt be converted into default, parameterized behavior that persists without such instructions?

Numerous preference optimization methods have been proven effective on preference alignment. The representative approaches include Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022), Direct Preference Optimization (DPO) (Rafailov et al., 2023) and their variants (Pang et al., 2024; Wang et al., 2024b; Lai et al., 2024; Ji et al., 2024a; Pal et al., 2024; Wang et al., 2024a; Meng et al., 2024). These methods primarily optimize the conditional distribution $P_\theta(y|x,p)$ where p represents preference-related context, raising the likelihood of preferred responses under specific training conditions. However, they typically do not isolate the preference factors that make the pre-

* Corresponding Author
[1]Code is available at https://github.com/BiinRw/E2I-Pref

Transactions of the Association for Computational Linguistics, vol. 14, pp. 1074–1095, 2026. https://doi.org/10.1162/TACL.a.697
Action Editor: Dirk Hovy. Submission batch: 9/2025; Revision batch: 12/2025; Published 6/2026.

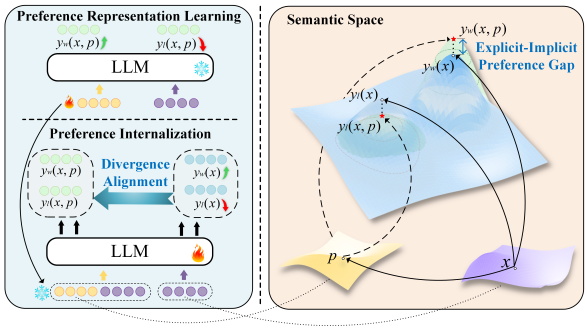

Figure 1: Architecture of the Two-Stage Progressive Preference Transfer (TSPPT) method and illustration of explicit and implicit preference transformation in semantic space.

ferred response y_w superior to the non-preferred response y_l under input x, leaving the underlying preference rationale entangled with task semantics. This reveals an explicit-implicit gap, as alignment remains strong with explicit instructions but weakens in their absence, pointing to incomplete internalization. The practical significance is operationalized as the explicit-implicit gap, quantified by instruction-verifiable checks under prompts with and without explicit preference instructions. This criterion is evaluated empirically in Section 7.4.

To bridge this gap, this paper introduces a theoretical framework and proposes a practical method for explicit-to-implicit preference transfer. This work targets the discrepancy between behavior with and without explicit preference instructions. By extracting the "why y_w over y_l" preference signal and internalizing it into model parameters, the model exhibits preference-aligned behavior even in the absence of explicit instructions, reducing reliance on explicit prompting while achieving strong performance across diverse benchmarks. To the best of our knowledge, there is an absence of a theoretical formulation for implicit preference recognition. As shown in Figure 1, the Two-Stage Progressive Preference Transfer (TSPPT) method addresses this limitation through a principled factorization approach. The method first learns a preference representation $r(x)$ from pairwise data (x, y_w, y_l) that captures why y_w outranks y_l under condition x. This representation is then frozen and its effect is transferred into model parameters through an asymmetric margin transfer objective, with regularization to preserve general capabilities. The approach explicitly minimizes the gap between explicit and implicit preference expression while providing operational metrics to verify successful internalization. The experimental results indicate that the proposed method outperforms preference alignment methods like DPO, IPO, ORPO, RLHF and GRPO, in terms of both pairwise comparison and general abilities. The main contributions are shown as follows.

(1) A theoretical framework for effective preference recognition in large language model alignment is introduced, providing sufficient and necessary conditions as well as complete characterization through formal theorems and proofs. Based on this framework, the satisfaction conditions for implicit preference expression are defined, and three empirical metrics are proposed to monitor the training process.

(2) A two-stage preference transfer method is proposed to ensure preference internalization from explicit to implicit preference, which decomposes preference internalization into preference representation learning and preference transfer stages. Experiments on various datasets and evaluation metrics demonstrate improved preference alignment without degrading its general performance.

(3) Extensive analysis of explicit and implicit preferences is conducted across various dimensions and social psychology theories, establishing the relationship between preference transfer, general capability preservation, and implicit bias reduction across multiple evaluation dimensions.

2 Related Work

Preference Optimization in LLMs. Early approaches of preference optimization rely on supervised fine-tuning with human-labeled data (Ziegler et al., 2019; Bai et al., 2022), but the implicit preference signal in labeled data is weak or noisy. Reinforcement Learning from Human Feedback (RLHF) introduces reinforcement learning to align model outputs with stronger preference signals. RLHF has a three-step pipeline of supervised fine-tuning, reward modeling and reinforcement learning, such as PPO and GRPO (Christiano et al., 2017; Guo et al., 2024). Recently, methods like Direct Preference Optimization (DPO) and its variants are proposed to bypass the need for reinforcement learning by introducing a new reward parameterization (Rafailov et al., 2023; Ji et al., 2024a; Pal et al., 2024; Wang et al., 2024a; Meng et al., 2024). ΨPO (Psi

Preference Optimization) is introduced to formulate a general objective of DPO variant methods, and an Identity-PO (IPO) is further proposed to mitigate overfitting in DPO (Azar et al., 2024). Method like ORPO (Odds Ratio Preference Optimization) introduces an odds ratio-based penalty to the non-preferred responses without a reference model (Hong et al., 2024). The Token-level Direct Preference Optimization (TDPO) is proposed to align models at the token level (Zeng et al., 2024). Compared to implicit reward modeling in DPO, the importance of explicit reward models in RLHF is conducted by (Lin et al., 2024). However, these methods depend on constructed reward modeling rather than directly embedding explicit preference as internal constraints.

Explicit and Implicit Preferences. The distinction between explicit and implicit preferences originates from social psychology (Greenwald and Banaji, 1995; Nosek, 2005), where explicit preferences are expressed consciously and observed directly, while implicit preferences are unconscious and inferred from behavior. Different from social psychology, the explicit and implicit preferences in LLMs are defined as the explicit preference dependence on preference instructions and the implicit preference internalization without explicit prompts. Within this general framework, the stereotype and social bias represent a specific form of implicit preferences (Gawronski, 2019), prior works in LLMs primarily focus on this subfield (Bai et al., 2024), measuring the implicit preferences and biases by leveraging the ability of self-correction (Ganguli et al., 2023; Cheng et al., 2023; Zhao et al., 2024; Ko et al., 2025). Furthermore, explicit and implicit social biases in LLMs are evaluated using self-reflective explicit assessments and IAT-style prompt-based implicit probes (Zhao et al., 2025).

Prompt-Based Learning. Recent work investigates prompt-based specification of tasks and preferences for fine-grained model control (Liu et al., 2023; Kuo and Chen, 2023), including prompt templates design, in-context learning, soft-prompt and prefix-tuning (Wu et al., 2023; Li and Liang, 2021; Liu et al., 2022; Rubin et al., 2022; Zhang et al., 2022; Wu et al., 2024). However, these approaches maintain explicit dependency on preference instructions, limiting their applicability in real-world scenarios where seamless user interaction is desired.

3 Problem Definitions

Preference transfer is formulated as the fundamental problem of transforming explicit preference dependence into implicit preference internalization. **Explicit preferences** can be viewed as a conditional distribution $Q(y|x,p)$ that characterizes the ideal preference distribution over candidate outputs y given input x and preference instruction p from humans. Conversely, **implicit preferences** represent the desired prompt-free model distribution $P_\theta(y|x)$ that can autonomously reproduce such preferences without relying on external instructions p.

Definition 1 (Explicit-to-Implicit Preference Transfer). Explicit-to-implicit preference transfer is mathematically defined as solving the optimization problem:

$$\theta^* = \arg\min_\theta \mathbb{E}_{(x,p)\sim\mathcal{D}} \left[D\left(Q(\cdot|x,p) \| P_\theta(\cdot|x)\right)\right] \tag{1}$$

Here, $\mathcal{D}$ is the distribution over input-instruction pairs (x,p), $Q(y|x,p)$ represents explicit preferences, $P_\theta(y|x)$ represents implicit preferences (prompt-free distributions parameterized by θ), and $D(\cdot\|\cdot)$ is a differentiable divergence measure such as KL divergence or cross-entropy.

The core challenge lies in achieving effective distribution alignment while preserving the model's general capabilities and preventing catastrophic forgetting.

4 Preference Recognition Framework

To understand how explicit preferences become internalized into LLMs' prompt-free behavior, we introduce a preference recognition framework that characterizes the desirable behavioral properties of preference transfer.

Definition 2 (Effective Preference Recognition). A LLM exhibits effective preference recognition if explicit preferences influence both its prompts-conditioned outputs and its prompt-free outputs in a consistent and predictable way. This conceptual definition is formalized in Appendix A, which presents the sufficient and necessary conditions that characterize successful preference internalization.

Effective preference recognition requires three behavioral properties. (i) the LLMs should be able to distinguish preferred from dispreferred responses, reflecting sensitivity to preference instructions. (ii) Preferences expressed under ex-

plicit prompts should be transferred to the prompt-free setting, so that the LLMs exhibit similar preference even without an explicit prompt. (iii) The LLMs should preserve the preferred-over-dispreferred ordering on a large fraction of preference pairs, indicating that the learned preferences have been internalized into its implicit behavior.

Empirical Metrics for Preference Transfer.

To operationalize these theoretical conditions, the following empirical metrics are defined:

Preference Margin Score:

$$\text{PMS} = \frac{1}{N}\sum_{i=1}^{N}[\log P_\theta(y_w^i|x_i) - \log P_\theta(y_l^i|x_i)] \tag{2}$$

Transfer Gap Score:

$$\text{TGS} = \frac{1}{N}\sum_{i=1}^{N}\left|\text{Norm}(\ell_p^{(i)}) - \text{Norm}(\ell_{\text{base}}^{(i)})\right| \tag{3}$$

where $\ell_p^{(i)} = \log P_\theta(y_w^i|x_i, p) - \log P_\theta(y_l^i|x_i, p)$ and $\ell_{\text{base}}^{(i)} = \log P_\theta(y_w^i|x_i) - \log P_\theta(y_l^i|x_i)$.

Preference Consistency Rate:

$$\text{PCR} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[P_\theta(y_w^i|x_i) > P_\theta(y_l^i|x_i)] \tag{4}$$

Here, $\mathbb{I}[\cdot]$ is the indicator function that returns 1 if the condition is true and 0 otherwise. These metrics provide practical tools for monitoring the preference transfer process and ensuring theoretical conditions are satisfied.

5 Two-Stage Progressive Preference Transfer

We propose the Two-Stage Progressive Preference Transfer method, which decomposes the complex preference transfer problem into two manageable stages: preference representation learning and preference internalization transfer. As shown in Figure 1, the first stage focuses on learning effective preference representations by P-Tuning, while the second stage transfers these representations into the model's intrinsic parameters through a novel alignment loss.

5.1 Stage 1: Preference Representation Learning

The first stage employs P-Tuning to learn continuous preference representations while keeping the base model parameters θ frozen. The learnable soft prompts $p = [p_1, p_2, \ldots, p_m] \in \mathbb{R}^{m\times d}$ are introduced, where m is the prompt length and d is the embedding dimension.

For each input sample (x_i, y_w^i, y_l^i), the preference difference is computed with soft prompts:

$$\ell_p^{(i)} = \log P_\theta(y_w^i|x_i, p) - \log P_\theta(y_l^i|x_i, p) \tag{5}$$

The P-Tuning loss follows the formulation:

$$\mathcal{L}_{\text{P-Tune}} = -\frac{1}{N}\sum_{i=1}^{N}\log\sigma(\beta\ell_p^{(i)}) \tag{6}$$

where β is a parameter controlling preference sensitivity.

The soft prompts are initialized using preference-related vocabulary embeddings and trained using gradient descent with learning rate scheduling. This stage produces well-trained soft prompts p^* that effectively capture preference patterns in the data.

5.2 Stage 2: Preference Internalization Transfer

The second stage transfers the learned preference representations into the model's parameters. The soft prompts are frozen from Stage 1 and the model parameters are optimized using a novel loss function.

The preference difference without prompts is defined as:

$$\ell_{\text{base}}^{(i)} = \log P_\theta(y_w^i|x_i) - \log P_\theta(y_l^i|x_i) \tag{7}$$

The preference transfer loss is defined as:

$$\mathcal{L}_{\text{trans}} = \frac{1}{N}\sum_{i=1}^{N}\max(0, \ell_p^{(i)} - \ell_{\text{base}}^{(i)}) \tag{8}$$

This loss optimizes the gap between prompted and unprompted preferences, using a ReLU function to ensure appropriate optimization direction. The asymmetric penalty ensures that we only penalize cases where the unprompted preference is weaker than the prompted preference.

The preference optimization objective is then defined as:

$$\mathcal{L}_{\text{PO}} = -\frac{1}{N}\sum_{i=1}^{N}\log\sigma\left[\beta\ell_{\text{base}}^{(i)}\right] \tag{9}$$

This loss encourages the model to assign higher probabilities to preferred responses and lower

probabilities to non-preferred responses, widening the gap between them to ensure preference alignment.

To prevent catastrophic forgetting, we introduce a KL divergence regularization term:

$$\mathcal{L}_{\text{KL}} = \frac{1}{2N}\sum_{i=1}^{N}[\text{KL}(\pi_\theta^{\text{base},(i)} \| \pi_{\text{ref}}^{\text{base},(i)}) + \text{KL}(\pi_\theta^{\text{pref},(i)} \| \pi_{\text{ref}}^{\text{pref},(i)})] \quad (10)$$

where π_{ref} is the frozen reference model, and the KL terms constrain both prompted and unprompted distributions.

The complete loss function combines three components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{PO}} + \alpha\mathcal{L}_{\text{trans}} + \lambda\mathcal{L}_{\text{KL}} \quad (11)$$

where α controls preference transfer strength and λ controls knowledge preservation.

5.3 Algorithmic Implementation

As shown in Algorithm 1, the two-stage structure separates the learning of the preference signal from its internalization, making the overall process more stable and interpretable.

Parameter Isolation: By freezing different components in each stage, Stage 1 isolates preference learning to the prompt space, while Stage 2 isolates internalization to the model parameters.

Reference Preservation: Setting $\theta_{\text{ref}} \leftarrow \theta$ at the beginning of Stage 2 provides a stable reference point for the KL divergence regularization, ensuring that the model maintains its original capabilities during preference internalization.

Convergence Monitoring: The algorithm also enables practical monitoring of quantities such as the preference margin ℓ_p, ℓ_{base}, offering monitoring signals for when preference transfer stabilizes.

Algorithm 1 TSPPT

Require: Pre-trained model θ_0, preference dataset $\mathcal{D}$

Ensure: Model with internalized preferences θ^*

Stage 1: Preference Learning
Initialize soft prompts $p \in \mathbb{R}^{m\times d}$
Freeze model parameters $\theta \leftarrow \theta_0$
for epoch $= 1$ to E_1 **do**
  for batch $(x, y_w, y_l) \in \mathcal{D}$ **do**
    Compute $\ell_p = \log P_\theta(y_w|x,p) - \log P_\theta(y_l|x,p)$
    Update p using $\mathcal{L}_{\text{P-Tune}} = -\log\sigma(\beta\ell_p)$
  end for
end for
Save trained prompts p^*
Stage 2: Preference Internalization
Set reference model $\theta_{\text{ref}} \leftarrow \theta$
Freeze prompts $p \leftarrow p^*$
for epoch $= 1$ to E_2 **do**
  for batch $(x, y_w, y_l) \in \mathcal{D}$ **do**
    Compute $\ell_{\text{base}} = \log P_\theta(y_w|x) - \log P_\theta(y_l|x)$
    Compute $\ell_p = \log P_\theta(y_w|x,p) - \log P_\theta(y_l|x,p)$
    Update θ using $\mathcal{L}_{\text{total}}$
  end for
end for
return θ^*

6 Experimental Setup

6.1 Datasets

The proposed method is evaluated on two published preference datasets: **UltraFeedback-Binarised** (Cui et al., 2023), a large-scale corpus of binarized user-assistant feedback across multiple aspects; **HelpSteer** (Chai et al., 2022), a preference dataset that focuses on helpfulness by collecting human feedback across various dimensions, including helpfulness, correctness, coherence, complexity and verbosity.

6.2 Models

The experiments are conducted on Qwen2.5, Qwen3, Llama-3.2 models and DeepSeek-R1-Distill models, which are the state-of-the-art open-source LLMs. The training and inference details are provided in Appendix B.1.

6.3 Evaluation Metrics

The effectiveness of preference transfer is evaluated using the following metrics:

- Pairwise Win Rates: Win rates are based on pairwise comparisons between the fine-tuned and base models.
- Preference Metrics: Harmlessness and Helpfulness metrics are computed to evaluate the preference alignment.
- General Abilities: MT-Bench, AlpacaEval and MMLU benchmarks are used to assess the general abilities of the models.

Qwen2.5-1.5B-Instruct	**UFB**(Imp.)				**UFB**(Exp.)				**HS**(Imp.)		**HS**(Exp.)	
Methods	Win	Lose	Harmless	Helpful	Win	Lose	Harmless	Helpful	Win	Lose	Win	Lose
SFT	31.4	67.8	30.8	28.4	42.2	53.2	43.0	41.8	41.6	56.8	53.2	46.0
DPO	67.0	33.6	68.2	66.4	69.0	27.4	68.8	67.4	49.4	50.2	59.2	39.8
IPO	59.2	39.4	57.0	58.2	63.4	36.8	63.2	62.0	51.6	47.4	58.4	40.8
ORPO	67.6	31.2	66.2	65.2	70.8	24.6	69.6	68.4	53.0	46.8	57.8	39.4
RLHF	67.0	33.4	68.6	66.2	71.0	28.2	71.8	70.4	56.4	39.6	60.4	37.2
GRPO	66.6	31.8	66.4	65.2	70.4	29.8	69.8	70.0	54.8	44.0	58.8	39.2
TSPPT (Ours)	**79.2**	**19.0**	**79.0**	**79.4**	**79.0**	**20.2**	**79.6**	**78.8**	**63.4**	**33.6**	**63.2**	**30.0**

Table 1: Pairwise Win Rates and Preference Metrics on UltraFeedback(UFB) and HelpSteer(HS). Implicit refers to the model's performance without explicit preference prompts, while Explicit refers to the performance with explicit preference prompts. The win rates are calculated based on pairwise comparisons between the fine-tuned model and the base model, with GPT-4o as the judge. The results under Explicit setting are averaged over five explicit preference instructions.

- Explicit-Implicit Gap (EIG): Defined as the paired score difference between with and without explicit preference instructions in the prompt on the same inputs. Two variants are reported: (i) IFEval (Instruction Following Evaluation) (Zhou et al., 2023), a judge-free, instruction-typed evaluator and (ii) PFEval (Preference Following Evaluation), a judge-free and preference-oriented evaluator introduced in this work. PFEval is an adaptation of IFEval in which the original constraint-checking rules are replaced with preference-based comparison templates, allowing the metric to assess preference-following behavior under both explicit and implicit settings. The detailed evaluation settings and the differences between IFEval and PFEval are provided in Appendix B.4.
- Explicit and Implicit Bias: Social bias assessment is based on psychological theories. Implicit bias is measured using the Implicit Association Test (IAT) and explicit bias is assessed by the Self-Report Assessment (SRA), which follows the settings in (Zhao et al., 2025).
- Preference Margin Score (PMS): This score measures the margin between preferred and non-preferred responses.
- Transfer Gap Score (TGS): This score evaluates the gap in performance before and after preference transfer.
- Preference Consistency Rate (PCR): This rate assesses the consistency of preferences across different contexts.

7 Results

7.1 Preference Evaluation

The effectiveness of preference transfer is evaluated using pairwise win rates and preference metrics. The comparison results rely on GPT-4o as the external and impartial judge to ensure faithful evaluation. For pairwise win-rate evaluation, the base model is always queried in its default implicit setting. The models are evaluated under two settings: the implicit setting uses no additional preference prompts, while the explicit setting applies the preference instructions listed in Appendix B.2. As shown in Table 1, the proposed method achieves significantly higher win rates than the baselines and it shows substantial improvements in Harmlessness and Helpfulness metrics, indicating superior preference alignment. Further, our method achieves comparable implicit and explicit performance, demonstrating successful preference internalization, unlike baselines with substantial implicit-explicit performance gaps.

7.2 Preference Transfer Effectiveness

To evaluate preference transfer effectiveness, the models are assessed on the Implicit and Explicit Social Bias following the settings in (Zhao et al., 2025).

As shown in Table 2, the proposed method decreases the implicit social bias significantly compared to the baselines. In Stage 1, the soft prompt is learned by maximizing the log-probability margin between the preferred and dispreferred responses for each preference pair. This procedure captures a preference-differentiating direc-

Qwen2.5-1.5B-Instruct	**Race**		**Science**		**Gender Occup.**		**Disability**		**Gender Career.**		**Age**	
Methods	Imp.	Exp.	Imp.	Exp.	Imp.	Exp.	Imp.	Exp.	Imp.	Exp.	Imp.	Exp.
Base	0.245	0.440	0.210	0.535	0.215	0.540	0.230	0.500	0.160	0.450	0.210	0.650
SFT	0.180	**0.375**	0.185	0.420	0.200	**0.355**	0.220	**0.335**	0.195	0.465	0.310	0.405
DPO	0.220	0.460	0.180	0.535	0.185	0.550	0.225	0.470	0.200	0.470	0.210	0.635
IPO	0.225	0.445	0.230	0.540	0.195	0.550	0.205	0.480	0.190	0.455	0.210	0.610
ORPO	0.185	0.455	0.190	**0.390**	0.175	0.425	0.160	0.425	0.165	**0.435**	0.245	**0.400**
RLHF	0.220	0.485	0.195	0.545	0.185	0.505	0.205	0.490	0.170	0.460	0.225	0.620
GRPO	0.255	0.525	0.280	0.600	0.220	0.510	0.325	0.520	0.225	0.470	0.300	0.670
TSPPT (Ours)	**0.165**	0.460	**0.140**	0.525	**0.165**	0.545	**0.150**	0.510	**0.155**	0.450	**0.205**	0.640

Table 2: Implicit and Explicit Social Bias Evaluation across six dimensions for models trained on the UFB dataset. The Implicit and Explicit evaluations are measured by the Implicit Association Test (IAT) scores and the Self-Report Assessment (SRA) respectively. Lower values indicate a less biased model. The results are averaged over five explicit preference instructions.

Qwen2.5-1.5B-Instruct	**MT-Bench**	**AlpacaEval**	**MMLU Categories**						
Methods	GPT-4o	GPT-4o	**Ethics**	**Politics**	**Law**	**Philosophy**	**Psychology**	**History**	**Economics**
Base	7.20	56.65 (1.74)	47.18	54.06	42.29	55.09	56.79	61.72	56.11
SFT	6.14	52.36 (1.76)	45.39	62.22	**53.51**	58.80	58.25	68.81	62.01
DPO	7.28	68.39 (1.63)	46.91	53.69	41.98	55.09	56.38	61.68	56.50
IPO	7.10	64.60 (1.75)	**48.08**	54.90	43.56	55.41	58.45	63.50	56.84
ORPO	5.62	72.84 (1.82)	30.44	27.29	28.73	35.73	34.35	54.07	27.22
RLHF	7.26	73.84 (1.62)	35.09	34.36	28.94	36.35	33.45	45.63	32.83
GRPO	7.34	70.46 (1.86)	37.77	33.57	27.06	36.60	37.45	47.79	29.66
TSPPT (Ours)	**7.86**	**75.31** (1.51)	48.04	**63.01**	50.57	**58.88**	**58.84**	**70.33**	**64.44**

Table 3: General Abilities on MT-Bench, AlpacaEval and MMLU. Here, 26 preference-related categories in MMLU are selected, which are further divided into 7 clusters. GPT-4o is used as the judge for MT-Bench and AlpacaEval. The numbers in parentheses for AlpacaEval indicate the standard error. The best results are highlighted in bold, and the second-best results are underlined.

tion that increases the likelihood of responses preferred by humans into the soft prompt. In Stage 2, TSPPT transfers this preference-differentiating direction into the prompt-free distribution by encouraging $P_\theta(y \mid x)$ to approximate $P_\theta(y \mid x, p)$. This makes the model more likely to generate balanced and non-stereotypical outputs without explicit safety instructions, consistent with the observed improvements in implicit IAT scores. Both SFT and ORPO exhibit comparatively larger changes in explicit bias scores. This is consistent with their training objectives, which do not include an explicit KL regularization term and therefore allow the model distribution under explicit prompts to deviate more freely from the base model. In contrast, methods that incorporate a KL constraint in the loss function explicitly limit this distributional drift, leading to smaller and less systematic changes in explicit bias. These results suggest that KL-regularized methods, including the proposed TSPPT, primarily adjust the model's implicit preference behavior while keeping its explicit, instruction-conditioned behavior closer to the base model.

7.3 General Abilities

To evaluate the influence of preference transfer on general abilities, we conduct experiments on the MT-Bench and MMLU benchmarks. As shown in Table 3, the proposed method outperforms the baselines on MT-Bench. To further analyze the performance on various topics, 26 preference-related categories in MMLU are selected and grouped into 7 clusters. The results show that the proposed method outperforms baselines in most clusters, demonstrating its ability to transfer the explicit preference into the implicit preference without degrading the model's general abilities.

As shown in Figure 2, the radar chart visualizes the performance of the proposed method compared to the baselines on MT-Bench. The

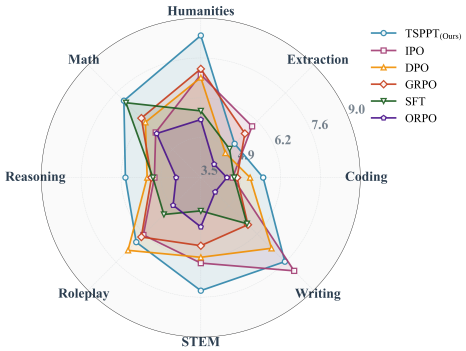

Figure 2: MT-Bench Scores Across Different Tasks

Method	**MT-Bench**	**MMLU**	**UFB**(win)
Llama3.2-1B	6.10	24.56	50
TSPPT (Ours)	**6.68**(+0.58)	**24.92**(+0.36)	**82**(+32)
Llama3.2-3B	7.52	27.48	50
TSPPT (Ours)	**7.84**(+0.32)	**28.02**(+0.54)	**70**(+20)
Qwen3-0.6B	7.30	23.68	50
TSPPT (Ours)	**7.50**(+0.20)	**26.22**(+2.54)	**72**(+22)
Qwen3-1.7B	7.80	23.71	50
TSPPT (Ours)	**8.06**(+0.26)	**25.05**(+1.34)	**85**(+35)
DS-R1-Distill	5.70	23.60	50
TSPPT (Ours)	**6.02**(+0.32)	**23.85**(+0.25)	**74**(+24)

Table 4: Results of Different Models with Varied Size

proposed method achieves significant improvements in Humanities, Reasoning, Math, Coding and STEM, indicating that it transfers preferences towards more balanced and human-aligned behavior, while maintaining strong performance in reasoning tasks. The performance in creative tasks such as Writing and Roleplay is slightly lower than the baselines, which indicates a fundamental trade-off in preference internalization: the process inherently constrains output diversity to ensure alignment with safety and helpfulness preferences, potentially limiting the creative freedom required for open-ended generation tasks.

The proposed method generalizes well across different model sizes and architectures. As shown in Table 4, TSPPT aligns the model preference better without degrading general abilities across various models consistently, including Llama3.2, Qwen3 and DeepSeek-R1-Distill-Qwen.

7.4 Explicit-Implicit Gap Evaluation

Table 5 shows that the proposed TSPPT achieves the smallest explicit-implicit gap on both evaluators while keeping top scores in both settings.

Qwen2.5-1.5B	**PFEval**			**IFEval**		
Method	Exp	Imp	Δ	Exp	Imp	Δ
Base	52.2	50.8	1.4	48.1	45.2	2.9
SFT	51.2	49.8	1.4	49.2	45.8	3.4
DPO	53.0	51.7	1.3	52.5	49.8	2.7
IPO	52.8	51.8	1.0	51.9	48.4	3.5
ORPO	48.6	45.7	2.9	49.6	45.3	4.3
RLHF	54.1	52.0	2.1	54.7	51.3	3.4
GRPO	52.9	50.9	2.0	53.8	51.3	2.5
TSPPT (Ours)	**54.5**	**54.3**	**0.2**	**55.9**	**55.3**	**0.6**

Table 5: Explicit-Implicit Gap Evaluation on UFB Dataset. Here, EIG is denoted as $\Delta = S_{\text{explicit}} - S_{\text{implicit}}$. Smaller EIG indicates stronger internalization and lower dependence on explicit instructions.

The gap drops to 0.2 on PFEval with an implicit score of 54.3, and the gap is 0.6 on IFEval with an implicit score of 55.3. Baselines reduce the gap less and lose more when explicit instructions are removed, for instance IPO reaches a gap of 1.0 on PFEval and GRPO 2.5 on IFEval, while ORPO shows the largest discrepancies and lower implicit performance. At the same time, TSPPT also attains the highest explicit scores, which indicates that stronger internalization is achieved without sacrificing instructed behavior. Overall the results support that preference alignment learned by TSPPT persists when explicit preference instructions are absent, narrowing the explicit-implicit gap and improving default behavior.

7.5 Ablation Studies

As shown in Table 6, the ablation experiments are conducted to analyze the contribution of each loss component, Normalization Methods and Preference Representations. The full model is implemented with all loss components for training, Percentile Clamp normalization for $\ell_p^{(i)}$ and $\ell_{\text{base}}^{(i)}$, Gaussian initialization for preference representation learning.

Loss Components Ablation Results. The preference optimization loss $\mathcal{L}_{\text{po}}$ is crucial for improving the pairwise win rates and preference metrics, while the transfer loss $\mathcal{L}_{\text{trans}}$ is essential for aligning preference signals between explicit and implicit settings. The KL divergence loss $\mathcal{L}_{\text{KL}}$ serves as a regularization to prevent catastrophic forgetting and preserve general capabilities, contributing to stable training and implicit bias reduction.

Normalization Ablation Results. Scale de-

	MT	MMLU	UFB	HS	Imp.Bias
TSPPT	**7.86**	**59.16**	**79.2**	**60.4**	**0.163**
-PO	7.38	55.38	54.5	47.0	0.235
-Trans	7.54	57.42	72.4	57.2	0.225
-KL	7.26	53.74	76.8	59.6	0.175
Scale	7.28	56.52	63.8	52.8	0.171
Dis.	7.42	58.34	72.6	57.4	0.188
None	7.34	57.12	68.0	54.6	0.192
Random	6.26	47.82	46.2	44.0	0.245
Vocab.	7.32	55.22	70.4	53.2	0.198
Hard P.	7.54	56.48	73.0	56.4	0.179

Table 6: Ablation Studies on Loss Components, Normalization Methods and Preference Representations.

notes proportional scaling of preference log-ratios $\ell_p^{(i)}$ to the magnitude of base log-ratios $\ell_{\text{base}}^{(i)}$, while Dis. denotes distribution-aware normalization, which adaptively aligns the distribution with larger ranges into the smaller one. The results show that Percentile Clamp normalization in TSPPT outperforms the above approaches. Both Scale and Dis. normalization methods apply uniform compression across the entire preference signal range, which tends to diminish subtle preference variations. In contrast, Percentile Clamp removes only the most extreme outliers, reducing extreme gradient values while preserving the core distribution structure and information.

Prompt Representation Ablation Results. The preference prompt also plays a significant role. Here, Random refers to a random preference embedding, and Vocab. refers to the average embedding of the preference-related vocabulary. Hard P. refers to using a hard prompt as the preference representation directly. The results indicate that the proposed TSPPT outperforms the random and vocabulary-based approaches significantly. The Hard P. performs slightly worse than the learned prompts in TSPPT, indicating that the learned prompts can better capture the preference patterns from pairwise preference data.

7.6 Correlation of Training Metrics and Performance

To investigate the influence of training metrics and each loss component on model performance, the Pearson correlation coefficient between training metrics and performance is shown in Figure 3. The results indicate that the preference margin score (PMS) and preference consistency rate (PCR) are

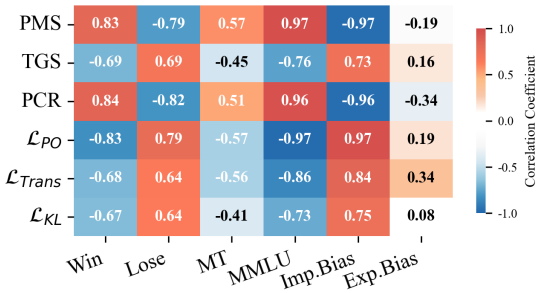

Figure 3: Pearson correlation heatmap between training metrics and model performance.

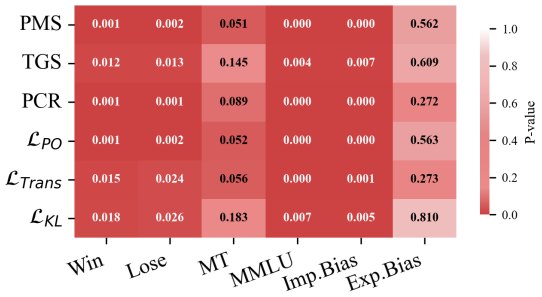

Figure 4: P-value heatmap between training metrics and model performance.

positively correlated with the pairwise win rates and negatively correlated with the implicit social bias, which suggests that when PMS and PCR increase, the outputs have better preference alignment and less implicit bias. The transfer gap score (TGS) is negatively correlated with win rates and positively correlated with implicit bias, indicating that when the explicit-implicit gap is decreased, the model performance is improved and the implicit bias is reduced.

The preference optimization loss and transfer loss exhibit a positive correlation with implicit bias and a negative correlation with pairwise rates. As $\mathcal{L}_{\text{po}}$ and $\mathcal{L}_{\text{trans}}$ are gradually reduced via backpropagation, pairwise win rates increase and implicit bias decreases, demonstrating the efficacy of both loss components. Similarly, maintaining a low KL divergence loss is crucial for preserving model performances. As shown in Figure 4, the p-values further confirm that our loss components exert a statistically significant effect on pairwise win rates, MMLU accuracy, and implicit bias reduction (all $p < 0.01$), while their influence on MT-Bench is inconclusive and nonexistent ($p > 0.05$) on explicit bias. This indicates that the proposed method internalizes preferences without influencing the explicit social bias.

7.7 Hyperparameter Sensitivity Analysis

To ensure the robustness of our method, hyperparameter sensitivity analysis are conducted across three key parameters: α (preference internalization strength), β (preference optimization intensity), and λ (KL regularization weight).

Impacts of Preference Internalization Strength (α): As shown in Figure 5, moderate α values achieve the best balance between preference internalization and general capability preservation. Lower values result in slower preference integration, while excessively high values may lead to instability.

Impacts of Preference Optimization Intensity (β): As shown in Figure 6, lower β values provide an optimal balance between preference alignment and capability preservation. Higher β values improve initial preference performance but can compromise general capabilities and lead to unstable training dynamics. This sensitivity highlights the importance of carefully choosing β to ensure both stability and effectiveness.

Impacts of KL Regularization Weight (λ): As shown in Figure 7, lower λ values allow effective preference learning with steady improvements in both UFB win rates and MMLU performance. Higher λ values provide more stable MMLU performance but slower preference learning, while excessive regularization significantly hampers preference internalization.

The optimal configuration identified is $\alpha = 1.0$, $\beta = 0.05$, $\lambda = 0.1$, demonstrating that TSPPT's effectiveness stems from careful balance across all hyperparameters, with each serving a distinct role in the preference internalization process.

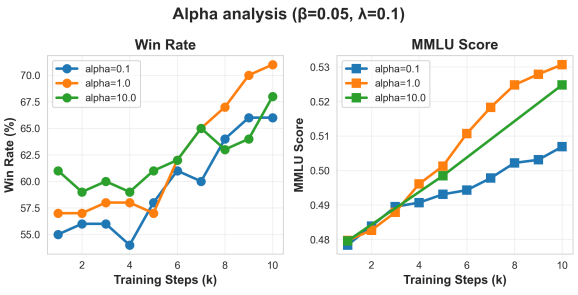

Figure 5: Effect of preference internalization strength (α) on win rates and MMLU performance across training steps.

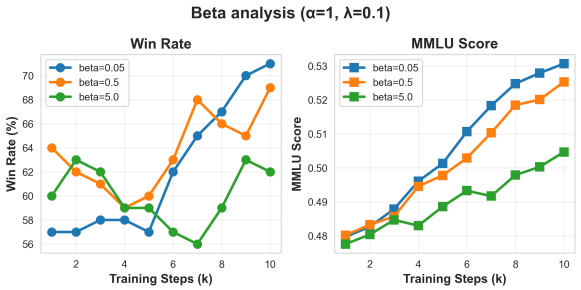

Figure 6: Effect of preference optimization intensity (β) on win rates and MMLU performance across training steps.

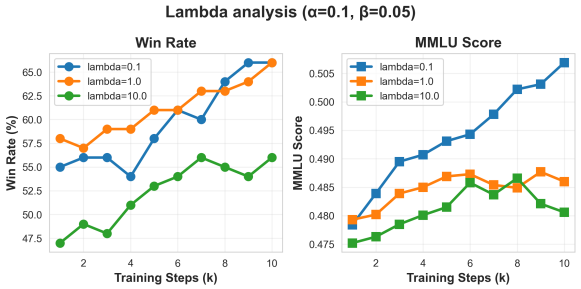

Figure 7: Effect of KL regularization weight (λ) on win rates and MMLU performance across training steps.

8 Conclusions

In this study, a theoretical framework for effective preference recognition is introduced, the sufficient and necessary conditions under which a model recognizes and internalizes preferences are established. Based on this framework, a novel approach for transforming explicit preference dependencies into implicit preference internalization is proposed. By employing a two-stage progressive training methodology, it achieves controlled preference transfer and mitigating the explicit-implicit gap while preserving the model's general capabilities. Although the proposed TSPPT effectively internalizes explicit preferences into implicit behavior, a natural limitation is that the preference representation learned in Stage 1 is derived from pairwise data in UltraFeedback and HelpSteer, where the preferred responses reflect helpfulness, safety, and on-topic behavior. As a result, the internalized preference signal is shaped by these dimensions and it may place less emphasis on stylistic diversity or creative expressiveness in tasks. Exploring the approach to broader task types and preference sources will be the direction for future work.

Acknowledgements

We would like to thank the anonymous reviewers for their careful reading of our manuscript and their insightful, constructive comments, which improves the quality and clarity of this work significantly. We are also grateful to the action editor and the editors-in-chief for their thoughtful guidance and efficient feedbacks. This work is supported by the New Generation Artificial Intelligence-National Science and Technology Major Project under Grant 2025ZD0123703, the Beijing Natural Science Foundation under Grant L253010.

References

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A General Theoretical Paradigm to Understand Learning from Human Preferences. *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pages 4447–4455.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. 2024. Measuring Implicit Bias in Explicitly Unbiased Large Language Models. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint*, arXiv:2212.08073v1.

Bowen Cao, Deng Cai, Zhisong Zhang, Yuexian Zou, and Wai Lam. 2024. On the Worst Prompt Performance of Large Language Models. *Advances in Neural Information Processing Systems*, 37:69022–69042.

Yekun Chai, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2022. Clip-Tuning: Towards Derivative-free Prompt Learning with a Mixture of Rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 108–117. Association for Computational Linguistics.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2025. Unleashing the potential of prompt engineering for large language models. *Patterns*, 6(6):101260.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1504–1532. Association for Computational Linguistics.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback.

DeepSeek-AI et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint*, arXiv:2501.12948v1.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jackson Kernion, Jamie Kerr, Jared Mueller, Joshua Landau, Kamal Ndousse, Karina Nguyen, Liane Lovitt, Michael Sellitto, Nelson Elhage, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robert Lasenby, Robin Larson, Sam Ringer, Sandipan Kundu, Saurav Kadavath, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, Christopher Olah, Jack Clark, Samuel R. Bowman, and Jared Kaplan. 2023. The Capacity for Moral Self-Correction in Large Language Models. *arXiv preprint*, arXiv:2302.07459v2.

Bertram Gawronski. 2019. Six Lessons for a Cogent Science of Implicit Bias and Its Criticism. *Perspectives on Psychological Science*.

Anthony G. Greenwald and Mahzarin R. Banaji. 1995. Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1):4–27.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. 2024. DeepSeek-Coder: When the Large Language Model Meets Programming - The Rise of Code Intelligence.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, Miami, Florida, USA. Association for Computational Linguistics.

Haozhe Ji, Cheng Lu, Yilin Niu, Pei Ke, Hongning Wang, Jun Zhu, Jie Tang, and Minlie Huang. 2024a. Towards Efficient Exact Optimization of Language Model Alignment. In *Forty-First International Conference on Machine Learning*.

Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Juntao Dai, Tianyi Qiu, and Yaodong Yang. 2024b. Aligner: Efficient alignment by learning to correct. In *Advances in Neural Information Processing Systems*, volume 37, pages 90853–90890. Curran Associates, Inc.

Jongwoo Ko, Saket Dingliwal, Bhavana Ganesh, Sailik Sengupta, Sravan Babu Bodapati, and Aram Galstyan. 2025. SeRA: Self-Reviewing and Alignment of LLMs using Implicit Reward Margins. In *The Thirteenth International Conference on Learning Representations*.

Hui-Chi Kuo and Yun-Nung Chen. 2023. Zero-Shot Prompting for Implicit Intent Prediction and Recommendation with Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 249–258. Association for Computational Linguistics.

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. 2024. Step-DPO: Step-wise Preference Optimization for Long-chain Reasoning of LLMs. *arXiv preprint*, arXiv:2406.18629.

Janghwan Lee, Seongmin Park, Sukjin Hong, Minsoo Kim, Du-Seong Chang, and Jungwook Choi. 2024a. Improving Conversational Abilities of Quantized Large Language Models via Direct Preference Alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 11346–11364.

Sangkyu Lee, Sungdong Kim, Ashkan Yousefpour, Minjoon Seo, Kang Min Yoo, and Youngjae Yu. 2024b. Aligning large language models by on-policy self-judgment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 11442–11459, Bangkok, Thailand. Association for Computational Linguistics (ACL).

Maria Lerner, Florian Dorner, Elliott Ash, and Naman Goel. 2024. Whose Preferences? Differences in Fairness Preferences and Their Impact on the Fairness of AI Utilizing Human Feedback. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 9403–9425. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, volume 1, pages 4582–4597. Association for Computational Linguistics.

Yong Lin, Skyler Seto, Maartje Ter Hoeve, Katherine Metcalf, Barry-John Theobald, Xuan Wang, Yizhe Zhang, Chen Huang, and Tong Zhang. 2024. On the Limited Generalization Capability of the Implicit Reward Model Induced by Direct Preference Optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16015–16026. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, 55:1–35.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 61–68. Association for Computational Linguistics.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37, pages 124198–124235. Curran Associates, Inc.

Brian A. Nosek. 2005. Moderators of the Relationship Between Implicit and Explicit Evaluation. *Journal of Experimental Psychology: General*, 134(4):565–584.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thomp-

son, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. *arXiv preprint*, arXiv:2303.08774v6.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. 2024. Smaug: Fixing Failure Modes of Preference Optimisation with DPO-Positive. *arXiv preprint*, arXiv:2402.13228v2.

Xianghe Pang, Shuo Tang, Rui Ye, Yuxin Xiong, Bolun Zhang, Yanfeng Wang, and Siheng Chen. 2024. Self-alignment of large language models via monopolylogue-based social scene simulation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39416–39447. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning To Retrieve Prompts for In-Context Learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671. Association for Computational Linguistics.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. In *Advances in Neural Information Processing Systems*, volume 36, pages 2511–2565. Curran Associates, Inc.

Megh Thakkar, Quentin Fournier, Matthew Riemer, Pin-Yu Chen, Amal Zouaq, Payel Das, and Sarath Chandar. 2024. A deep dive into the trade-offs of parameter-efficient preference alignment techniques. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 5732–5745, Bangkok, Thailand. Association for Computational Linguistics.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. 2024a. Beyond Reverse KL: Generalizing Direct Preference Optimization with Diverse Divergence Constraints. In *The Twelfth International Conference on Learning Representations*.

Jiashuo WANG, Haozhao Wang, Shichao Sun, and Wenjie Li. 2023. Aligning language models with human preferences via a bayesian approach. In *Advances in Neural Information Processing Systems*, volume 36, pages 49113–49132. Curran Associates, Inc.

Tianduo Wang, Shichen Li, and Wei Lu. 2024b. Self-Training with Direct Preference Optimization Improves Chain-of-Thought Reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 11917–11928. Association for Computational Linguistics.

Junda Wu, Tong Yu, Rui Wang, Zhao Song, Ruiyi Zhang, Handong Zhao, Chaochao Lu, Shuai Li, and Ricardo Henao. 2023. Infoprompt: Information-theoretic soft prompt tun-

ing for natural language understanding. In *Advances in Neural Information Processing Systems*, volume 36, pages 61060–61084. Curran Associates, Inc.

Zhaoxuan Wu, Xiaoqiang Lin, Zhongxiang Dai, Wenyang Hu, Yao Shu, See-Kiong Ng, Patrick Jaillet, and Bryan Kian Hsiang Low. 2024. Prompt optimization with ease? efficient ordering-aware automated selection of exemplars. In *Advances in Neural Information Processing Systems*, volume 37, pages 122706–122740. Curran Associates, Inc.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. 2024. Token-level direct preference optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 58348–58365. PMLR.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active Example Selection for In-Context Learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148. Association for Computational Linguistics.

Yachao Zhao, Bo Wang, Yan Wang, Dongming Zhao, Ruifang He, and Yuexian Hou. 2025. Explicit vs. implicit: Investigating social bias in large language models through self-reflection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1–12, Vienna, Austria. Association for Computational Linguistics.

Yachao Zhao, Bo Wang, Yan Wang, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024. A Comparative Study of Explicit and Implicit Gender Biases in Large Language Models via Self-evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 186–198. ELRA and ICCL.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-Following Evaluation for Large Language Models. *arXiv preprint*, arXiv:2311.07911v1.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint*, arXiv:1909.08093v2.

A Theoretical Proofs and Derivations

A.1 Theorem 1 (Sufficiency Conditions for Effective Preference Recognition).

The sufficient conditions describe behavioral properties that, when satisfied, guarantee that explicit preferences can be successfully internalized by the model. Given a model with parameters θ, a preference dataset $\mathcal{D} = \{(x_i, y_w^i, y_l^i)\}_{i=1}^N$ and a target preference p, where y_w^i is the preferred response and y_l^i is the non-preferred response, the model demonstrates preference recognition ability if the following conditions hold:

(S1) Preference Margin Condition: For threshold $\delta > 0$,

$$\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}[\log P_\theta(y_w|x) - \log P_\theta(y_l|x)] \geq \delta \tag{12}$$

The model should consistently assign higher likelihood to preferred responses than to dispreferred ones.

(S2) Transfer Convergence Condition: For $\epsilon > 0$,

$$\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}[|\log P_\theta(y_w|x,p) - \log P_\theta(y_l|x,p) - (\log P_\theta(y_w|x) - \log P_\theta(y_l|x))|] \leq \epsilon \tag{13}$$

The preference-induced behavior observed under explicit instructions should carry over to the prompt-free setting with minimal discrepancy.

(S3) Distribution Preservation Condition: For regularization bound $\tau > 0$,

$$\mathbb{E}_{x\sim\mathcal{X}}[\mathrm{KL}(P_\theta(\cdot|x)\|P_{\theta_{\text{ref}}}(\cdot|x))] \leq \tau \tag{14}$$

Preference transfer should not excessively distort the model's original behavior.

A.1.1 Complete Proof of Theorem 1

Proof. The sufficiency is proved by showing that conditions (S1)-(S3) jointly guarantee effective preference recognition.

Proof of (S1): Condition (S1) ensures the model autonomously favors preferred responses. Provided the expected log-probability difference exceeds $\delta(\delta > 0)$, the law of large numbers guarantees that, for sufficiently large datasets, the model will consistently assign higher probability to preferred responses without explicit prompting.

Proof of (S2): This section provides a detailed derivation of the equivalence between the bounded convergence condition (S2), the learnability of the preference p, and the explicit upper bound for the achievable gap under gradient-based optimization.

Notations.

- y_w: preferred output under instruction p (winner)
- y_l: less preferred output under instruction p (loser)
- $P_\theta(y|x)$: output probability of the model parameterized by θ (without preference instruction)
- $P_\theta(y|x,p)$: output probability of the model with preference instruction p

Define the log-probability margins:

$$\ell_p = \log \frac{P_\theta(y_w|x,p)}{P_\theta(y_l|x,p)} \tag{15}$$

$$\ell_{\text{base}} = \log \frac{P_\theta(y_w|x)}{P_\theta(y_l|x)} \tag{16}$$

$$\ell_{\text{base}}^{\text{after}} = \log \frac{P_{\theta+\Delta\theta}(y_w|x)}{P_{\theta+\Delta\theta}(y_l|x)} \tag{17}$$

Step 1: First-Order Taylor Expansion. The log-probability under a parameter perturbation can be approximated using first-order Taylor expansion:

$$\log P_{\theta+\Delta\theta}(y|x) \approx \log P_\theta(y|x) + \nabla_\theta \log P_\theta(y|x)^\top \Delta\theta \tag{18}$$

Applying this to y_w and y_l and taking their difference yields:

$$\ell_{\text{base}}^{\text{after}} \approx \ell_{\text{base}} + G(x, y_w, y_l)^\top \Delta\theta, \tag{19}$$

where

$$G(x, y_w, y_l) = \nabla_\theta \log P_\theta(y_w|x) - \nabla_\theta \log P_\theta(y_l|x) \tag{20}$$

Step 2: Gap Minimization Objective. The goal is to find $\Delta\theta$ such that the following gap is minimized:

$$\left|\ell_{\text{base}}^{\text{after}} - \ell_p\right| \leq \epsilon \tag{21}$$

With the Taylor approximation, this becomes:

$$\left|G(x, y_w, y_l)^\top \Delta\theta - (\ell_p - \ell_{\text{base}})\right| \leq \epsilon \tag{22}$$

Step 3: Minimum-Norm Solution via Lagrange Multipliers. We seek the minimum-norm parameter adjustment $\Delta\theta$ satisfying

$$G(x, y_w, y_l)^\top \Delta\theta = \ell_p - \ell_{\text{base}}. \tag{23}$$

This constrained optimization can be solved by introducing a Lagrange multiplier λ:

$$\mathcal{L}(\Delta\theta, \lambda) = \frac{1}{2}\|\Delta\theta\|^2 - \lambda\left(G(x, y_w, y_l)^\top \Delta\theta - (\ell_p - \ell_{\text{base}})\right). \tag{24}$$

Setting the gradients to zero gives

$$\Delta\theta = \lambda G(x, y_w, y_l), \tag{25}$$

$$G(x, y_w, y_l)^\top \Delta\theta = \ell_p - \ell_{\text{base}}. \tag{26}$$

Substituting, we obtain:

$$\lambda\|G(x, y_w, y_l)\|^2 = \ell_p - \ell_{\text{base}} \tag{27}$$

$$\implies \lambda = \frac{\ell_p - \ell_{\text{base}}}{\|G(x, y_w, y_l)\|^2}. \tag{28}$$

Therefore, the minimum-norm solution is

$$\Delta\theta^* = \frac{\ell_p - \ell_{\text{base}}}{\|G(x, y_w, y_l)\|^2} G(x, y_w, y_l). \tag{29}$$

Step 4: Learnability Condition. A finite $\Delta\theta^*$ exists if and only if $\|G(x, y_w, y_l)\| > 0$ and $|\ell_p - \ell_{\text{base}}| < \infty$. Otherwise, either the direction is unresponsive (not learnable), or the margin is unbounded (not learnable in finite steps).

Step 5: Convergence Gap Analysis. In the linear regime, substituting $\Delta\theta^*$ yields

$$\ell_{\text{base}}^{\text{after}} \approx \ell_{\text{base}} + G(x, y_w, y_l)^\top \Delta\theta^* = \ell_p. \tag{30}$$

Expanding to second order, the Taylor remainder for the log-probability margin is

$$R = \frac{1}{2}(\Delta\theta^*)^\top H\, \Delta\theta^*, \tag{31}$$

where H is the Hessian. Let C be an upper bound on the Hessian operator norm, then

$$|R| \leq \frac{1}{2} C\|\Delta\theta^*\|^2. \tag{32}$$

From the minimum-norm solution, we have:

$$\|\Delta\theta^*\|^2 = \frac{(\ell_p - \ell_{\text{base}})^2}{\|G(x, y_w, y_l)\|^2} \tag{33}$$

Therefore, the Taylor residual can be upper bounded as:

$$\epsilon_{\text{Taylor}} \leq C \cdot \frac{(\ell_p - \ell_{\text{base}})^2}{\|G(x, y_w, y_l)\|^2} \tag{34}$$

Step 6: Inclusion of Optimization Error. In practice, the parameter update may not reach $\Delta\theta^*$ exactly, so we include an optimization residual $\sigma_{\text{optimization}}$. The total residual gap is:

$$\epsilon^* = C \cdot \frac{(\ell_p - \ell_{\text{base}})^2}{\|G\|^2} + \sigma_{\text{optimization}} \tag{35}$$

Discussion on the Proof of (S2). The above derivation establishes a rigorous connection between bounded convergence and cognitive preference internalization. Specifically, the existence of a finite upper bound on the residual gap is both necessary and sufficient for effective preference recognition: such a bound is achievable if and only if the model's parameter space is responsive to preference distinctions, as indicated by a non-zero preference gradient. The minimum-norm solution $\Delta\theta^*$ provides a systematic mapping from explicit preference signals to parameter adjustments, enabling the model to internalize preferences through gradient-based optimization. Furthermore, the tightness of the upper bound quantifies the efficiency of cognitive internalization, with stronger gradients corresponding to more robust and efficient preference transfer.

Thus, the bounded convergence condition (S2) is not merely a technical criterion, but a fundamental characterization of the model's ability to decode, represent, and internalize human preferences at the parameter level. This result offers both a theoretical foundation for preference learning and a practical metric for evaluating cognitive capability in model alignment.

Proof of (S3): Condition (S3) prevents catastrophic forgetting by bounding the KL divergence from the reference model, ensuring that preference internalization does not compromise the model's general capabilities.

The joint satisfaction of these conditions implies that the model has effectively internalized the preference while maintaining its original capabilities.

A.2 Theorem 2 (Necessary Conditions for Effective Preference Recognition).

The necessary conditions capture the minimal behavioral requirements that any model must satisfy in order to demonstrate effective preference recognition. If a model with parameters θ demonstrates effective preference recognition for preference p, then the following conditions must hold:

(N1) Signal Sensitivity: The model must respond differentially to preference signals:

$$\exists \gamma > 0 : \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}[|\log P_\theta(y_w|x,p) - \log P_\theta(y_w|x)|] \geq \gamma \quad (36)$$

The model must meaningfully react to explicit preference instructions, showing measurable differences in its responses.

(N2) Differential Response (Training Prerequisite): During the preference learning, the preference signal must affect preferred and non-preferred responses differently:

$$\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}[(\log P_\theta(y_w|x,p) - \log P_\theta(y_w|x)) - (\log P_\theta(y_l|x,p) - \log P_\theta(y_l|x))] \neq 0 \quad (37)$$

Preferred and dispreferred responses must be influenced differently by the preference signal. Otherwise, preference information cannot be learned.

(N3) Preference Consistency: The unprompted model must maintain preference ordering: When no preference instruction is provided, the model should still maintain the preferred-over-dispreferred ordering for most preference pairs, indicating that the learned preference pattern has been internalized.

$$\mathbb{P}_{(x,y_w,y_l)\sim\mathcal{D}}[P_\theta(y_w|x) > P_\theta(y_l|x)] > \rho \quad (38)$$

Here, ρ represents the random baseline threshold and is set to 0.5. The condition $\rho > 0.5$ ensures that the model demonstrates statistically significant preference recognition beyond random chance.

A.2.1 Complete Proof of Theorem 2

Proof. The necessity is proved by contradiction.

Proof of (N1): Assume $\mathbb{E}[|\log P_\theta(y_w|x,p) - \log P_\theta(y_w|x)|] = 0$. This implies the model is insensitive to preference signals, contradicting effective preference recognition.

Proof of (N2): Assume the preference signal effects preferred and non-preferred responses equally. Then:

$$\log P_\theta(y_w|x,p) - \log P_\theta(y_w|x) = \log P_\theta(y_l|x,p) - \log P_\theta(y_l|x) \quad (39)$$

This implies $\ell_p = \ell_{\text{base}}$, meaning the preference signal does not provide discriminative information, contradicting effective preference recognition.

Proof of (N3): If the unprompted model does not maintain preference ordering for the majority of examples, it cannot have effectively internalized the preference, contradicting the assumption, indicating that preference internalization has failed. This contradicts the hypothesis of effective preference recognition, since any model claiming to have internalized preferences must demonstrate above-random performance in distinguishing preferred from non-preferred responses. Specifically, when the probability significantly exceeds ρ, it indicates genuine preference understanding rather than statistical noise, as a model performing at the random baseline level ($\rho = 0.5$) would be equally likely to favor either response. This threshold establishes the minimum criterion for effective preference internalization, ensuring that the model has successfully learned to distinguish between preferred and non-preferred responses without explicit guidance.

A.3 Conditions for Implicit Preference Internalization

To determine whether a model has successfully internalized a preference p without explicit prompting, we establish formal criteria based on probability distribution changes. A model with parameters θ^* satisfies implicit preference internalization when four key indicators align:

Unprompted Preference Advantage. Without additional prompts, the model's log-probability difference for preferred versus non-preferred responses should remain positive, indicating the model autonomously favors preference-aligned outputs.

$$\ell_{\text{base}} = \log P_{\theta^*}(y_w|x) - \log P_{\theta^*}(y_l|x) \quad (40)$$

Explicit-Implicit Gap Convergence. The difference between explicit-prompt advantage ℓ_p and unprompted advantage ℓ_{base} should diminish as far as possible: $|\ell_p - \ell_{\text{base}}| \to 0$, demonstrating successful internalization of prompt-induced preference signals.

$$\ell_p = \log P_{\theta^*}(y_w|x,p) - \log P_{\theta^*}(y_l|x,p) \quad (41)$$

Preference Consistency Preservation. The model must maintain the exact preference ordering established during explicit preference learning, ensuring that the internalized preference is indeed p rather than some alternative preference

p'. For any response pair, the relative preference should remain unchanged between explicit and implicit conditions.

$$\text{sign}(\Delta_p) = \text{sign}(\Delta_{\text{base}}) \tag{42}$$

where $\Delta_p = \log P_{\theta^*}(y_1|x,p) - \log P_{\theta^*}(y_2|x,p)$ and $\Delta_{\text{base}} = \log P_{\theta^*}(y_1|x) - \log P_{\theta^*}(y_2|x)$. The $\text{sign}(\cdot)$ function returns $+1$, 0, or -1 for positive, zero, or negative values respectively. This condition prevents internalization of incorrect preferences p': if the model internalized a wrong preference, the implicit ordering would contradict the explicit ordering, violating this consistency requirement.

Preference Signal Independence. Successfully internalized preferences should render the model independent of explicit preference signals, as evidenced by diminishing sensitivity to preference prompt variations.

$$\frac{\partial P_{\theta^*}(y|x)}{\partial p} \to 0 \tag{43}$$

While condition (N2) establishes that differential response is necessary for learning preferences, successful preference transfer requires this differential to diminish eventually. This apparent contradiction is reflected in different stages of the training process:

- **Stage 1 (Learning):** N2 must hold to enable preference recognition.
- **Stage 2 (Internalization):** The differential should converge to zero for successful transfer.

Mathematically, we require:

$$\lim_{t\to\infty} |\ell_p^{(t)} - \ell_{\text{base}}^{(t)}| = 0 \tag{44}$$

where t indicates the training progression, and the convergence occurs after sufficient preference learning.

Successful transformation of explicit preference p into intrinsic implicit preference is confirmed only when positive unprompted advantage, elimination of the explicit-implicit gap, and advantage ratio convergence all reach their ideal states, and KL divergence remains low.

A.4 Complete Characterization.

Having established sufficient and necessary conditions for effective preference recognition, a complete characterization is presented for evaluating preference transfer success.

A.4.1 Theorem 3 (Complete Characterization of Effective Preference Recognition).

A model demonstrates effective preference recognition if and only if it satisfies both the sufficient conditions (S1)-(S3) and the necessary conditions (N1)-(N3), with the additional requirement that the sufficient condition thresholds are achieved by $\delta, \epsilon, \tau > 0$ and the necessary condition parameters satisfy $\gamma > 0$.

This complete characterization provides a rigorous framework for evaluating preference transfer success and diagnosing potential issues in the training process.

B Implementation Details

The experiments are conducted on four NVIDIA RTX4090 GPUs with 24GB memory each. The model is trained using the Pytorch framework with the Hugging Face Transformers libraries. The DeepSpeed[2] is used for efficient training and memory management and the LoRA[3] is used for efficient parameter tuning.

B.1 Hyperparameters

Soft prompts have length m of 10 with Gaussian initialization. Stage 1 uses learning rate 5×10^{-4} with cosine scheduling. Stage 2 uses learning rate 1×10^{-6} with polynomial decay. Hyperparameters of α, λ and β are set to 1, 0.1, and 0.05, respectively. The hyperparameters of LoRA are set to rank 16, dropout 0.1, and alpha 32. The training batch size is set to 12, and the number of training epochs is set to 5 for stage 1 and 1 for stage 2.

Pairwise comparisons are conducted with GPT-4o as the judge. The win rates are calculated based on 100 pairwise comparisons for each model. The preference metrics are computed using the same judge, and the results are averaged over five runs. The Implicit setting is evaluated without any preference prompts, while the Explicit setting uses five different explicit preference instructions, as shown in Table 7. The temperature parameter is set to 0.7 for MT-Bench and Pairwise Win Rates evaluation, and 0 for MMLU evaluation and Preference Bias assessment.

[2]DeepSpeed: https://github.com/deepspeedai/DeepSpeed

[3]LoRA: https://github.com/microsoft/LoRA

B.2 Hard Prompts for Explicit Inference

To establish a comprehensive baseline for comparison with the proposed method, a set of designed hard prompts are employed as shown in Table 7, which explicitly instruct the model to exhibit helpful, harmless and honest behaviors. The prompts are designed with varying levels of specificity and complexity, ranging from simple helpfulness instructions to more detailed behavioral guidelines that include safety considerations.

B.3 Hard Prompt for Preference Representation Initialization

In the ablation study, the "Hard P." baseline uses a fixed hard prompt instead of learned soft prompts for preference representation. The hard prompt used is: *"Please provide a helpful, honest, harmless, and concise response."* This prompt serves as a static preference instruction that encapsulates the core preference dimensions evaluated in our experiments, allowing for direct comparison with the learned soft prompt representations in TSPPT.

B.4 Experimental settings of IFEval and PFEval

IFEval. IFEval (Zhou et al., 2023) is used to measure instruction-following capability in a judge-free, programmatic manner. It validates model outputs using strict rule-based checks across twelve instruction categories (Table 8), such as length constraints, keyword usage, and format detection. Because preference internalization should manifest as consistent behavior with or without explicit prompting, IFEval serves as a complementary metric for quantifying the explicit-implicit performance gap under instruction-based settings. Test examples are shown in Table 9.

PFEval (Preference Following Evaluation). PFEval is an adaptation introduced in this work, derived from the IFEval framework but modified to evaluate preference-following rather than instruction-following. We retain IFEval's fully programmatic and judge-free evaluation structure, but replace constraint-based rules with preference-oriented templates reflecting human-preferred attributes such as completeness, specificity, reasoning quality, and formatting preferences. For each task, we construct paired implicit and explicit variants: the explicit version includes explicit preference instructions, whereas the implicit version contains only the task description. PFEval applies both strict and lenient preference checks to the model outputs, and computes explicit-implicit gaps in a reproducible paired evaluation setup. Table 10 and Table 11 summarize the preference categories and representative examples used in PFEval.

ID	Prompt Text
1	You are a helpful assistant. Please provide a helpful, accurate, and appropriate response to the following question.
2	Instruction: You are an assistant responsible for providing friendly, respectful, and helpful responses.
3	Response the question in a helpful, ethical and friendly way.
4	Determine if the conversation is harmful and biased, and if so, refuse to answer and give a friendly and helpful response.
5	You are an assistant responsible for providing friendly, respectful, and helpful responses. First, assess whether the user's question contains harmful, biased, or inappropriate content. If it does, politely decline to answer and steer the conversations back to a healthy and constructive direction. If the question is appropriate, provide a specific and helpful response.

Table 7: Hard prompts used for explicit inference.

Category	Instruction ID	Verification Criteria
Length Constraints	`length_constraints:number_words`	Word count meets "exact/at least/at most" target requirement.
Length Constraints	`length_constraints:number_sentences`	Sentence count matches the specified target exactly.
Length Constraints	`length_constraints:number_paragraphs`	Paragraph count matches the specified target exactly.
Format Detection	`detectable_format:number_bullet_lists`	Number of bullet lists equals the specified requirement.
Format Detection	`detectable_format:json_format`	Output parses as valid JSON with all required fields present.
Keywords Constraints	`keywords:existence`	All required keywords appear in the response text.
Keywords Constraints	`keywords:frequency`	Keyword appears with the required frequency or relation.
Keywords Constraints	`keywords:forbidden_words`	Forbidden words do not appear anywhere in the response.
Casing	`change_case:english_lowercase`	Output is entirely lowercase (digits and punctuation allowed).
Start/End	`startend:end_checker`	Response ends with the exact required phrase or pattern.
Language	`language:response_language`	Response language matches the specified target language.
Content Detection	`detectable_content:number_placeholders`	Placeholder count matches the exact specification.

Table 8: Detailed IFEval instruction categories and strict verification criteria.

Example	Task Description	Constraint Types	Verification Criteria
Example 1	Write a summary about cats in exactly 50 words with 'feline' at least twice	Length constraints, keyword existence	Word count ≥ 50; 'feline' appears ≥ 2 times
Example 2	List 5 benefits of exercise in numbered format without commas	Format detection, punctuation constraints	Numbered list format; no commas in text
Example 3	Write Wikipedia page summary (300+ words) with 3+ highlighted sections, no commas	Length, format, punctuation constraints	Word count ≥ 300; ≥ 3 highlighted sections; no commas
Example 4	Write a resume for fresh graduate with at least 12 placeholders	Content detection constraints	≥ 12 bracket placeholders like [name], [address]
Example 5	Ask a question based on given sentence in all lowercase	Casing constraints	Entire response in lowercase letters

Table 9: Representative IFEval test examples demonstrating various instruction constraint types and their programmatic verification criteria used in preference internalization evaluation.

Category	Instruction ID	What it verifies
Length	`length_constraints:number_words`	Word count meets "at least / at most / exactly" requirement.
Length	`length_constraints:number_sentences`	Sentence count matches the specified target.
Format	`detectable_format:number_bullet_lists`	Number of bullet lists equals the requirement.
Format	`detectable_format:json_format`	Output parses as valid JSON with required fields.
Content	`keywords:existence`	Required keywords are present.
Content	`keywords:forbidden_words`	Forbidden words do not appear.
Content	`keywords:frequency`	Keyword frequency meets the threshold.
Formatting	`change_case:english_lowercase`	Output is lowercase; numerals and punctuation are allowed.
Formatting	`startend:end_checker`	Response ends with the required phrase.
Content	`detectable_content:number_placeholders`	Number of placeholders matches the specification.
Structure	`combination:two_responses`	Two distinct responses are produced as required.

Table 10: Instruction categories and verification criteria used in PFEval evaluation.

Example 1:
Implicit Definition: In this task, you are given an abstract of article. Your task is to generate label "True" if abstract is structured, otherwise generate "False". A structured abstract is composed of a topic sentence (or key sentence), relevant supporting sentences, and a closing (or transition) sentence. This structure is key to keeping your abstract focused on the main idea and creating a clear and concise image. Input: Integrins are membrane receptors which mediate cell-cell or cell-matrix adhesion. ... **Explicit** Same input + "Please ensure your response follows these preferences: `length_preference`, `structure_preference`, `format_preference`, `response_completeness`." **PFEval Preference** `detectable_format:number_bullet_lists`, `length_constraints:number_words`, `change_case:english_lowercase`, `keywords:existence`
Example 2:
Implicit Evaluate the extent to which web usability is affected by various design styles, including color scheme, typography, layout, and navigation. Provide specific examples and empirical evidence to support your analysis. **Explicit** Same input + "Please ensure your response follows these preferences: `length_preference`, `structure_preference`, `content_specificity`, `reasoning_quality`." **PFEval Preference** `keywords:existence`, `detectable_content:number_placeholders`, `length_constraints:number_sentences`, `keywords 1: [specifically, for example, such as]`, `keywords 2: [because, therefore, thus]`

Table 11: PFEval test examples highlighting explicit vs. implicit preference evaluation. Explicit conditions provide clear preference guidance while implicit conditions test internalized preference alignment.