

WorldPM: Scaling Human Preference Modeling via Real-World Feedback

Binghai Wang^{1,2}, Runji Lin², Keming Lu², Le Yu², Zhenru Zhang², Fei Huang^{2*},
Chujie Zheng², Kai Dang², Yang Fan², An Yang², Binyuan Hui², Dayiheng Liu², Tao Gui^{1*},
Qi Zhang¹, Xuanjing Huang¹, Yu-Gang Jiang¹, Bowen Yu^{2*}, Jingren Zhou², Junyang Lin²

¹ Fudan University ² Qwen Team, Alibaba Group

Abstract

Inspired by the success of scaling language modeling through massive real-world corpora, we demonstrate that similar power-law patterns exist in preference modeling. We propose **World Preference Modeling (WorldPM)**, which leverages naturally occurring signals from real-world user interactions as a scalable source of preference data. We curate a 15M-scale preference dataset from online forum and conduct large-scale experiments on models from 1.5B to 72B parameters. Across 10 objective preference test sets with consensus-based labels, we observe two distinct scaling patterns: (1) common-sense preference (e.g., identifying factual errors) exhibits stable log-linear improvement across all model sizes, while (2) domain preference (e.g., coding, math, knowledge QA) exhibits a capacity-gated pattern where only sufficiently large models achieve meaningful scaling. Through training dynamics analysis, we reveal the learning mechanism behind preference scaling: the model first learns to discriminate using surface preferences, but counter-surface examples progressively compel it to discover non-surface preferences as training scales. Further experiments validate WorldPM as an effective foundation for preference fine-tuning, broadly improving generalization across human preference datasets of varying sizes (7K, 100K, and 800K), with gains exceeding 5% on many key subtasks.

1 Introduction

Kaplan et al. (2020) demonstrate that language modeling performance exhibits scaling laws: by pre-training on massive real-world corpora, model capabilities are continuously pushed to new frontiers (Touvron et al., 2023; Qwen et al., 2024). We argue that similar scaling potential exists in preference modeling, and thus propose **World Preference Modeling (WorldPM)** to unlock it. WorldPM uses naturally occurring aggregate signals from large-scale community interactions as a scalable preference-modeling foundation.

To capture this real-world preference consensus at scale, online forums provide a natural solution. As platforms that facilitate large-scale interaction and opinion sharing, online forums aggregate vast amounts of user preference data through voting mechanisms, substantially advancing research on collective decision making in social sciences and psychology (Easley & Kleinberg, 2010; Pal et al., 2012; Hu et al., 2008). Despite the immense value of these preference signals, they have rarely been effectively leveraged at scale by artificial intelligence.

In this study, we compare preference data from StackExchange, Reddit, and Quora, then use StackExchange as the sole source of the 15M training pairs for every reported WorldPM model; Reddit and Quora appear only in source-selection and cross-source analyses. We conduct systematic experiments on models (Qwen et al., 2024) ranging from 1.5B to 72B parameters. Across 10 objective preference test sets spanning diverse domains, we consistently observe that test losses decrease steadily as training scale and model capacity increase. Further categorized by difficulty, this scaling trend manifests in two distinct patterns:

*Corresponding authors: feihu.hf@alibaba-inc.com, tgui@fudan.edu.cn, and yubowen.ybw@alibaba-inc.com.

(1) In common-sense preference evaluation, test losses decrease log-linearly with training scale. On basic tasks such as identifying responses containing factual errors, or those that are fluent but off-topic or incomplete, models of all sizes exhibit highly consistent power-law scaling. This suggests that large-scale training can substantially enhance reward models’ ability to discriminate response quality deficiencies (Park et al., 2024; Zeng et al., 2024).

(2) In domain preference evaluation, we observe a capacity-gated scaling pattern. On tasks requiring genuine domain knowledge (such as mathematics, code, and knowledge-based QA), smaller models (1.5B, 7B) show limited improvement, while only models of sufficient scale (32B, 72B) begin to exhibit log-linear decreases in test loss. This suggests that domain preference discrimination is an inherently challenging task that requires sufficient model capacity to benefit from large-scale weakly supervised data, echoing capacity-dependent phenomena observed in language modeling (Wei et al., 2022).

Through analysis of WorldPM’s training dynamics, we reveal the learning mechanism underlying preference modeling’s scalability. We discover an **asymmetric learning dynamic**: the model first exploits surface preferences (e.g., favoring longer responses) to classify the majority of examples, but the dataset also contains substantial counter-surface examples where shorter responses are preferred. These examples produce persistent losses that surface preferences cannot resolve, compelling the model to learn deeper semantic representations and internalize a robust consensus on human intrinsic utility (Geirhos et al., 2020).

Finally, we validate WorldPM’s strong generalization as a foundation for preference fine-tuning. We conduct comprehensive experiments across human preference datasets of varying sizes (7K, 100K, and 800K samples) and evaluate on 7 benchmarks with 16 subtasks. WorldPM consistently yields significant performance gains across all dataset scales, with improvements exceeding 5% on many key subtasks. Furthermore, performance gains show a clear positive correlation with pre-training scale, demonstrating WorldPM’s scalable advantages as a preference foundation model.

2 Background & Related Work

2.1 Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF) has emerged as a critical technique for aligning large language models (LLMs) with human preferences (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Grattafiori et al., 2024; Yang et al., 2024; Qwen et al., 2024; DeepSeek-AI et al., 2025). A central component is *reward modeling* (Bai et al., 2022; Zheng et al., 2023), which captures human preferences and is also referred to as preference modeling (PM) (Ouyang et al., 2022). We use both terms interchangeably in this work.

A preference model (PM) learns from pairwise comparison data (x, y_0, y_1) , where label $Y \in \{0, 1\}$ indicates which response is preferred. The model computes scores $r_\theta(x, y_0)$ and $r_\theta(x, y_1)$ for each response using reward function r_θ . Following the Bradley-Terry model (Bradley & Terry, 1952), the preference probability is:

$$P(Y = 0|x, y_0, y_1) = \text{sigmoid}(r_\theta(x, y_0) - r_\theta(x, y_1)), \quad (1)$$

The PM replaces the LLM’s last decoding layer with a linear layer to output a scalar score and is trained on preference dataset $\mathcal{D}$ to minimize the Bradley-Terry (BT) loss (Sun et al., 2024):

$$\mathcal{L}_{\text{BT}} = -\mathbb{E}_{(x, y_0, y_1, Y) \sim \mathcal{D}} [\log P(Y|x, y_0, y_1)], \quad (2)$$

The trained PM is used to evaluate policy quality. In this work, we adopt Best-of-N (BoN) sampling as a fine-tuning-free alignment evaluation method (Gao et al., 2023; Coste et al., 2023; Ivison et al., 2024), which selects the highest-scoring response from N candidates generated for each prompt.

2.2 Preference Modeling

Existing preference modeling and alignment methods can be primarily divided into two categories: The first category is scalar reward modeling based on the Bradley-Terry (BT) model. The second category involves text-based generative reward modeling methods like GenRM and LLM-as-judge (Liu et al., 2025b; Chen et al., 2025; Gu et al., 2026; Whitehouse et al., 2025). WorldPM focuses on the BT scalar model because its fixed scalar loss is defined identically across model and data scales, enabling controlled comparisons; this methodological choice is not a claim that BT is superior to generative reward models (GenRMs). Scaling analyses for GenRMs are left to future work. Previous research often utilized small-scale, high-quality human-annotated datasets (typically containing 10^3 – 10^5 examples) (Wang et al., 2025) or AI-annotated datasets such as UltraFeedback (Cui et al., 2024), which may exhibit systematic biases (Chiang & yi Lee, 2023; Fu & Liu, 2025). Recent work has also explored large-scale human-AI collaborative curation to improve both scale and quality simultaneously (Liu et al., 2025a). In contrast, WorldPM leverages large-scale forum data, which inherently contains human preference signals to empirically study scaling laws in reward modeling.

2.3 Preference Model Pre-training

Due to the high cost of preference annotation, Askell et al. (2021) propose using public forum data for Preference Model Pre-training (PMP). Building on their research, we contribute several novel findings beyond simply scaling data and models: (1) we discover two distinct scaling patterns: common-sense preference follows a stable power-law, while domain preference exhibits capacity-gated scaling at larger sizes; (2) we reveal an asymmetric learning mechanism where models initially exploit surface features before counter-surface examples drive them to learn intrinsic utility; and (3) we show that surface-feature interference fundamentally limits subjective evaluations for studying scaling laws. Additionally, we conduct preference fine-tuning on larger, diverse datasets across rich benchmarks, extending well beyond their limited 5K-sample, in-distribution experiments.

3 Modeling World Preference

3.1 Data Source Selection

Data Collection: To model world preference, we collect data from multiple public forums, including StackExchange (a professional Q&A platform), Reddit (a social news and community discussion platform), and Quora (a knowledge sharing community). These forums contain posts with user responses that receive upvotes and downvotes, naturally creating preference patterns. For each prompt x (post), we sample two responses with different net votes (upvotes minus downvotes) to form preference pairs (y_0, y_1) , where the label Y points to the response with higher votes.

Data Selection: Through extensive training and validation across different data sources, StackExchange emerges as the highest-quality source, achieving the best generalization both across external out-of-domain test sets and across other data sources (Table 2). Notably, StackExchange also demonstrates the strongest cross-source generalization ability, transferring well to Reddit, Quora, and even human-annotated HelpSteer2 (Figure 6). We therefore select StackExchange as the sole source for the 15M WorldPM training set; Reddit and Quora are retained only for source-selection and cross-source analyses. Furthermore, we verify that our findings are robust to data construction choices: vote score differences between preference pairs do not significantly affect performance (Table 3), and training on different topic domains within StackExchange yields consistent results across all benchmarks (Table 4), suggesting that the preference patterns captured are domain-agnostic. Detailed analyses of different data sources can be found in Appendix A.1. Additional details on forum data analysis, preference pair sampling strategies (including vote margin controls) and training examples are provided in Appendices A.2, A.3 and F.

3.2 Experimental Setup

Training Methods: We follow the standard preference modeling framework (Section 2.1). All models are trained using a learning rate of $3e-6$ and a 10K batch size, with ablation studies provided in Appendix B.1.

Evaluation Methods: Preference test sets from RM benchmarks generally fall into **objective** and **subjective** categories. Objective test sets have consensus-based labels (e.g., factual correctness or code functionality), while subjective ones reflect individual annotator preferences. We focus on objective test sets because current subjective benchmarks conflate surface-feature confounds, pluralistic values, and potentially non-transitive structures, which obstruct a single scaling signal (Section 3.5). Within objective test sets, we further distinguish by difficulty: (1) **Common-sense preference** identifying basic flaws such as factual errors or off-topic answers, requiring no specialized knowledge, and (2) **Domain preference** requiring domain expertise, such as mathematical proofs or code correctness.

Specifically, we evaluate WorldPM on diverse objective test sets from multiple RM benchmarks: PPE (Frick et al., 2024) (MMLU-Pro, IFEval, GPQA, MATH, and MBPP-Plus (Liu et al., 2023)), RMB (Zhou et al., 2024) (Helpfulness), RM-Bench (Liu et al., 2024) (chat, code, math), RewardBench (Lambert et al., 2025) (chat-hard, reasoning), and Offset Bias (Park et al., 2024). We uniformly use their preference pairs to compute test losses (i.e., BT loss) for consistent evaluation. Detailed descriptions of these benchmarks and our evaluation methods are provided in Appendices B.2 and B.3. For the six domain benchmarks supporting the capacity-gated claim, we run a 13-gram overlap screen against all 15M training pairs. Manual review finds no confirmed benchmark-derived training pair among the automatically flagged items (Appendix A.5).

3.3 Scaling Laws

We collect as many posts, responses, and votes as possible from StackExchange to construct preference pairs, yielding approximately 15M pairs, and conduct world preference modeling on Qwen2.5 models (Qwen et al., 2024) ranging from 1.5B to 72B parameters. Subsequently, we evaluate WorldPM’s out-of-domain generalization on various objective test sets using BT loss (Equation 2). We find that WorldPM exhibits clear but distinct scaling patterns on common-sense and domain preferences (Figure 1).

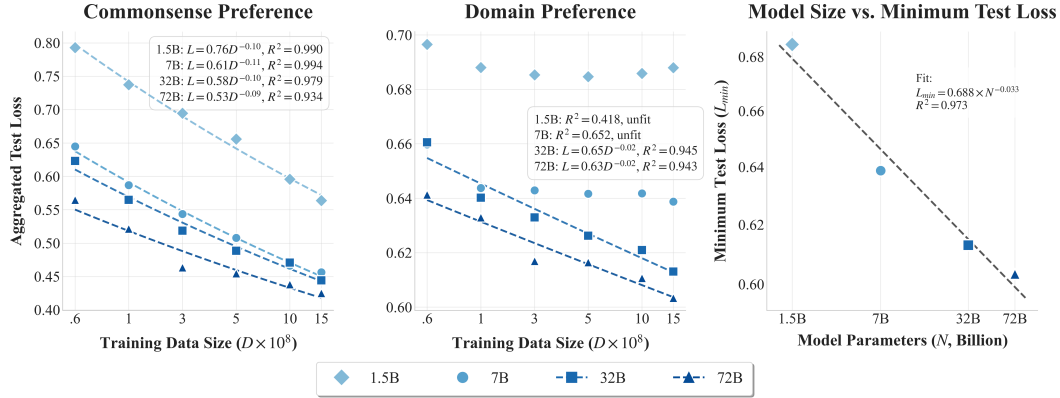


Figure 1: Scaling patterns of WorldPM. Left: Test loss on common-sense preference evaluation follows a power law with training data size ($R^2 > 0.9$). Middle: Domain preference evaluation shows a capacity-gated scaling pattern, where power-law decrease is observed only at 32B+ scales while smaller models show limited improvement. Right: Minimum test loss across four model scales shows a decreasing trend consistent with a power-law fit, though the limited number of data points warrants cautious interpretation.

Common-sense preference exhibits stable power-law scaling. Our common-sense preference evaluation comprises three test sets: factual error detection (from RM-Bench chat),

OffsetBias, and LLMBar (from RewardBench chat-hard). These tasks test basic judgments on which humans broadly agree, such as identifying factual errors or distinguishing fluent but off-topic answers. As shown in Figure 1 (left), all model sizes (1.5B to 72B) exhibit consistent power-law decrease with aggregated R^2 exceeding 0.9. Prior studies (Park et al., 2024; Zeng et al., 2024) have identified PM vulnerabilities that RLHF can exploit (e.g., favoring superficially good but unfaithful responses). Our findings suggest these vulnerabilities may stem from limited training data.

Domain preference exhibits capacity-gated scaling. We evaluate domain preference using six objective test sets: coding (MBPP-Plus, HumanEvalPack), mathematics (MATH), knowledge-based QA (MMLU-Pro, GPQA), and instruction following (IFEval). As shown in Figure 1 (middle), small models (1.5B, 7B) produce nearly flat scaling curves with low goodness of fit (0.418 and 0.652, respectively), while large models (32B, 72B) exhibit clear power-law decrease. This capacity-gated pattern suggests that **effective domain preference scaling requires a minimum model capacity**, though we note that this observation is based on four model sizes within a single architecture family and warrants further validation.

Preliminary evidence for parameter scaling. Figure 1 (right) plots minimum test loss against model parameters. A power-law fit ($L_{\min} = 0.688 \times N^{-0.033}$, $R^2 = 0.973$) suggests that performance improves with model size. However, this observation relies on only four data points from a single model family, limiting the statistical reliability of the fit. We treat this as a suggestive trend; validation across more model scales and architectures is needed before drawing conclusions about parameter scaling laws in preference modeling.

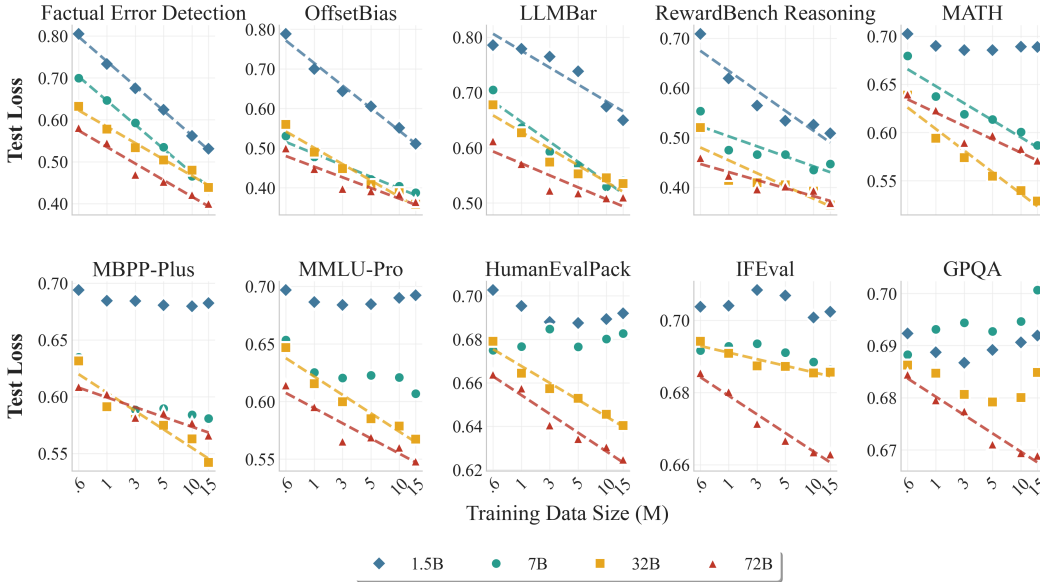
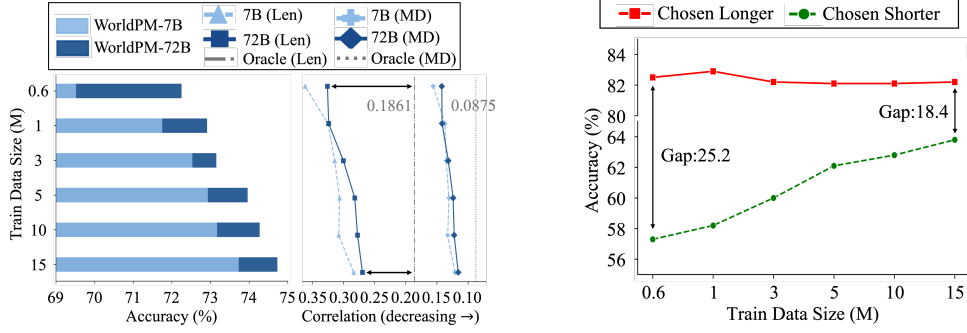


Figure 2: **Task-wise scaling analysis across ten objective test sets.** Larger models demonstrate broader generalization, maintaining scaling even on the most challenging tasks (e.g., GPQA) where smaller models fail to benefit from additional training data.

Figure 2 further reveals that **scaling in preference modeling is capacity-gated**. Across the ten objective test sets, smaller models progressively lose the ability to benefit from additional training data on more demanding tasks: 1.5B flattens first (e.g., MBPP-Plus, MATH), followed by 7B and 32B. In contrast, only the largest model (72B) maintains meaningful scaling on the most challenging benchmarks (e.g., GPQA). This demonstrates that larger models can extract effective preference signals and scale across a significantly broader range of tasks.



(a) Left: validation accuracy improves continuously with training and model scale. Right: correlation between model predictions and surface features (length, markdown formatting) is high early in training but decreases with training scale and model size, indicating that the model naturally reduces its reliance on surface features through large-scale training.

(b) Asymmetric learning dynamics. The model rapidly fits majority-surface-pattern data (Chosen Longer) to 82% accuracy while gradually incorporating minority counter-surface instances (Chosen Shorter, starting at only 57%), with the latter showing logarithmic growth.

3.4 Learning Mechanism of Preference Scaling

The above scaling laws show that preference modeling performance improves continuously with training scale and model size. But what does the model actually learn? Prior research has documented that preference models tend to learn surface-level shortcuts (Geirhos et al., 2020; Holtzman et al., 2021), such as preferring longer or more richly formatted responses rather than truly understanding content quality. This length and style bias (Dubois et al., 2024; Chiang et al., 2024) remains a central challenge. We investigate whether **large-scale training can naturally overcome this surface feature dependence** by analyzing the relationship between model predictions and surface features during WorldPM training.

Large-scale training naturally overcomes surface preference bias. Given a preference pair (x, y_0, y_1) , let S , P , and G denote which response has better surface attributes, the model’s prediction, and the human label, respectively. We use the Phi coefficient (ϕ) (Cramér, 1946) to measure their correlations and examine two surface attributes: S_{length} (longer response) and $S_{markdown}$ (more markdown markers).

As shown in Figure 3a, validation accuracy improves continuously with training and model scale (left), while $\phi(P, S)$ steadily decreases toward the Oracle level $\phi(G, S)$ (right). Early in training, $\phi(P, S)$ far exceeds the Oracle, consistent with shortcut learning (Geirhos et al., 2020; Holtzman et al., 2021; Zhang et al., 2016). The joint trend of rising accuracy and declining $\phi(P, S)$ is consistent with the hypothesis that the model progressively learns to judge content quality rather than merely favoring longer or better-formatted responses.

Asymmetric learning dynamics reveal the mechanism behind scaling. We split the validation set by length into two groups: Chosen Longer (the preferred response is longer than the rejected one) and Chosen Shorter (vice versa). As shown in Figure 3b, the model starts at 82% accuracy on Chosen Longer but only 57% on Chosen Shorter; as training progresses, Chosen Longer remains stable while Chosen Shorter grows logarithmically. The model initially exploits the “longer is better” heuristic for strong majority performance, but the substantial counter-surface examples (where the preferred response is shorter) produce persistent losses that this heuristic cannot resolve, compelling the model to learn deeper semantic representations that go beyond surface preferences.

3.5 Limitations of Subjective Evaluation

The preceding analysis concerns objective test sets. We organize the difficulty of subjective scaling evaluation into three distinct sources. **(a) Surface-feature confounds:** labels correlate with response length and formatting; the style-control experiments below

directly quantify only this source. **(b) Pluralistic values:** different groups can hold coherent but incompatible preferences without a single ground truth (Sorensen et al., 2024). **(c) Non-transitive structure:** pairwise preferences need not admit a global scalar order (Gu et al., 2021; Duan et al., 2017), whereas BT represents each response by one score. These latter two possibilities are meaningful structure, not merely noise, and are discussed rather than measured here. The retained comparisons produced from aggregate StackExchange scores are acyclic by construction; this property of the constructed labels does not establish transitivity of individual or underlying population preferences (Appendix A.4).

3.5.1 Style Control Method

Inspired by Chiang et al. (2024), we separate style from content in preference model evaluation. For a preference pair (x, y_0, y_1) , let $r_\theta(x, y_0)$ and $r_\theta(x, y_1)$ be the model scores, with their score difference $D(x, y_0, y_1) = r_\theta(x, y_0) - r_\theta(x, y_1)$. Let $V \in \mathbb{R}^S$ be the style feature vector, where S is the number of features. We use four style features: token length, markdown lists, headers, and bold elements. The style difference is normalized as follows:

$$Z_i(x, y_0, y_1) = \text{normalize} \left(\frac{V_i(y_0) - V_i(y_1)}{V_i(y_0) + V_i(y_1)} \right), \quad i = 1, \dots, S. \quad (3)$$

The final evaluation metric combines score difference and style difference as $R = D\alpha + Z^T\beta$, where weights α and β are learned through linear regression with the preference model scores frozen, minimizing the binary cross-entropy loss:

$$\hat{\alpha}, \hat{\beta} = \arg \min_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^S} \frac{1}{n} \sum_{i=1}^n -(Y_i \log(\text{sigmoid}(R_i)) + (1 - Y_i) \log(1 - \text{sigmoid}(R_i))), \quad (4)$$

Explicitly modeling style factors maximizes the capture of style-related gains in evaluation outcomes across different models, isolating style influence when comparing models. Style control uses $R_i = D_i\alpha + Z_i^T\beta$ as the final score, while the no-control condition fixes $\beta = 0$.

3.5.2 Comparative Analysis of Subjective and Objective Test Sets

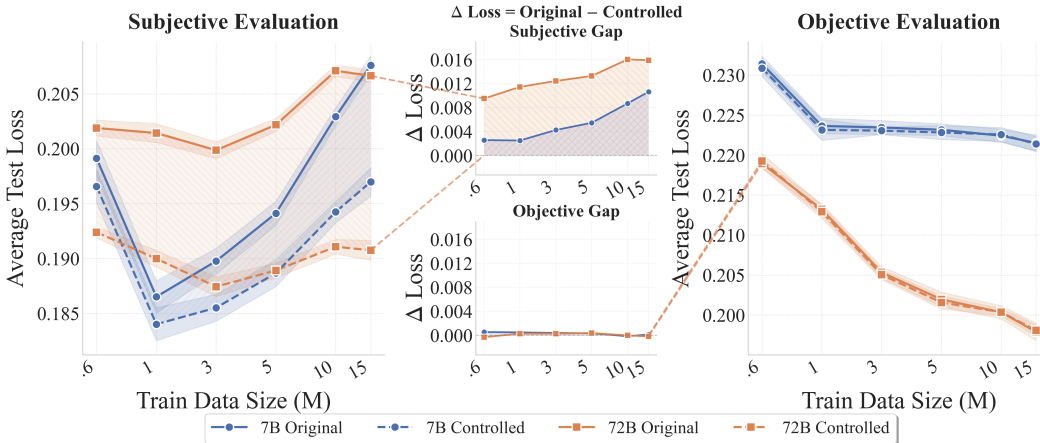


Figure 4: Test loss comparison on subjective tasks (left) and objective tasks (right) with and without style control. Subjective tasks are substantially affected by surface features, with this effect amplifying with scale; objective tasks are nearly unaffected.

As shown in Figure 4, the two types of test sets exhibit contrasting patterns. On **objective test sets** (right), the curves with and without style control nearly overlap, confirming that evaluation results genuinely reflect content discrimination ability. On **subjective test sets** (left), the uncontrolled curve shows a non-monotonic trend (first decreasing then increasing). After style control, the curve becomes noticeably smoother, indicating that this apparent behavior largely stems from spurious correlations between surface features and subjective

labels. **More critically, this interference continuously amplifies with training scale and model size**, as the gap between Original and Controlled curves keeps widening. This is a direct consequence of Section 3.4: as WorldPM progressively reduces its reliance on surface features, its predictions diverge further from the surface-correlated subjective labels, widening the gap.

Furthermore, length and markdown formatting are only a few quantifiable dimensions among surface features. Subjective preferences are also influenced by many latent factors (e.g., tone, level of detail, phrasing style) that cannot be easily separated or controlled. Even controlling for just a few quantifiable surface features fundamentally changes subjective scaling curves; the full spectrum of latent factors would only exacerbate this problem. Therefore, **the surface-feature confounds directly measured here increasingly affect the observed subjective scaling curves as model capability improves. Current benchmarks therefore require methods that separate such confounds from genuine preference diversity before they can support a single scaling-law characterization.**

4 Preference Fine-Tuning based on WorldPM

4.1 Experimental Setup

We evaluate WorldPM as an effective foundation for preference fine-tuning using three human preference datasets: HelpSteer2 (Wang et al., 2024b), UltraFeedback (Cui et al., 2024), and pair_data.v2 800K wsafety¹ from RLHFlow (hereafter referred to as RLHFlow). HelpSteer2 comprises approximately 7K preference comparisons across five dimensions. Since we consider only a single reward, we selected helpfulness scores as the preference labels. UltraFeedback contains four responses per question; we extracted approximately 100K preference pairs. RLHFlow encompasses approximately 800K data points. These datasets of varying sizes enable us to understand the effectiveness boundary of WorldPM. We use batch size 128 for HelpSteer2 and 512 for others, training each for at least two epochs. Models are selected based on validation loss.

We validate the effectiveness of WorldPM in preference fine-tuning through two approaches. First, we evaluate using various RM benchmark metrics (consistent with Section 3.2, including subjective, domain preference, and common-sense preference metrics). Following Section 3.5.1, we compute scores as preference prediction accuracy, where preference is determined by the sign of reward difference R (with or without style control). Second, we employ best-of- N sampling for alignment evaluation. The alignment performance is assessed on two benchmarks - Arena Hard and Alpaca Eval - following our previously described evaluation protocol. Specifically, we generate 64 samples from Qwen2.5-7B-Instruct and Qwen2.5-72B-Instruct for the 7B and 72B models.

4.2 Evaluation Results

Table 1 shows the results for 72B models with style-controlled scores (see Appendix C.3 for 7B and uncontrolled results, which show consistent findings). **WorldPM initialization improves performance across domains and fine-tuning datasets of varying scales.** Given the high cost of preference annotation, WorldPM provides a valuable foundation for PM training. Key findings include:

Subjective domains: PPE-Human and RMB-Helpful show notable improvements, suggesting WorldPM learns useful general representations for subjective preference. RewardBench-Chat shows slight decrease but remains above 97%, indicating metric saturation.

Domain preference domains: PPE’s domain preference metrics, RM-Bench’s code and math metrics, and RewardBench’s reasoning metrics demonstrate broad improvements.

Common-sense preference domains: Smaller datasets (e.g., HelpSteer2) show notable improvements, while larger ones exhibit minimal changes. WorldPM itself already achieves

¹https://huggingface.co/datasets/RLHFlow/pair_data.v2.80K_wsafety

Table 1: Effect of WorldPM initialization on 72B model’s PM fine-tuning performance across different evaluation categories. Abbreviations: RB - RewardBench, RM-B - RM-Bench.

Metrics	Helpsteer2		UltraFeedback		RLHFlow	
	w/o WorldPM	w/ WorldPM	w/o WorldPM	w/ WorldPM	w/o WorldPM	w/ WorldPM
Subjective Evaluation						
PPE-Human	64.76	66.68 (↑2.97%)	65.12	66.94 (↑2.80%)	68.14	68.93 (↑1.16%)
RMB-Helpful	77.76	78.53 (↑1.00%)	79.80	81.73 (↑2.41%)	79.26	80.89 (↑2.05%)
RB-Chat	98.04	98.88 (↑0.85%)	97.63	96.37 (↓1.29%)	98.46	97.91 (↓0.57%)
Domain Preference Evaluation						
PPE-MBPP	70.50	78.76 (↑11.72%)	75.02	79.46 (↑5.92%)	82.28	82.39 (↑0.14%)
PPE-GPQA	57.66	59.63 (↑3.42%)	59.22	61.19 (↑3.33%)	62.52	63.75 (↑1.97%)
PPE-IFEval	59.24	63.63 (↑7.42%)	60.57	64.34 (↑6.22%)	64.26	65.39 (↑1.76%)
PPE-MATH	69.94	77.17 (↑10.33%)	72.77	78.16 (↑7.41%)	79.36	83.01 (↑4.60%)
PPE-MMLU	70.20	74.96 (↑6.79%)	72.42	75.08 (↑3.67%)	79.04	79.30 (↑0.32%)
RM-B-Code	60.75	69.30 (↑14.08%)	67.32	70.18 (↑4.23%)	72.15	73.46 (↑1.82%)
RM-B-Math	72.78	77.69 (↑6.75%)	73.72	77.79 (↑5.51%)	87.81	86.20 (↓1.83%)
RB-Reason	96.65	98.25 (↑1.66%)	96.30	97.48 (↑1.23%)	97.52	98.15 (↑0.64%)
Common-sense Preference Evaluation						
RB-Chat-Hard	84.87	87.28 (↑2.84%)	84.54	83.00 (↓1.82%)	84.21	84.54 (↑0.39%)
RM-B-Chat	81.01	82.17 (↑1.44%)	80.23	79.46 (↓0.97%)	85.66	84.50 (↓1.36%)
OffsetBias	83.14	89.76 (↑7.96%)	87.29	88.05 (↑0.88%)	87.89	87.47 (↓0.47%)
Alignment Evaluation						
Arena Hard	88.52	89.72 (↑1.36%)	89.16	90.37 (↑1.36%)	90.41	90.82 (↑0.45%)
Alpaca Eval	60.34	63.39 (↑5.07%)	59.95	63.73 (↑6.31%)	65.51	66.26 (↑1.14%)
Arena Hard-SC	88.47	90.54 (↑2.34%)	88.82	90.74 (↑2.16%)	91.06	91.76 (↑0.77%)
Alpaca Eval-LC	53.26	55.45 (↑4.12%)	56.83	59.80 (↑5.23%)	56.95	55.58 (↓2.40%)

high accuracy on these metrics (around 90% for RM-Bench Chat and OffsetBias, as shown in Figure 18). The post-fine-tuning decrease shows WorldPM’s initial gains are gradually lost, approaching the performance of models without WorldPM initialization, indicating easily modifiable features.

Alignment evaluation: Most metrics improve with and without style control (-LC and -SC denote their respective default style control strategies). Exceptions in RLHFlow’s Alpaca Eval may be attributed to RLHFlow’s large scale and high quality, making WorldPM’s benefits less pronounced.

4.3 The Impact of WorldPM Training Scale on Fine-tuning

To study how WorldPM’s training scale affects fine-tuning, we experiment with different 72B WorldPM checkpoints (trained on 5M, 10M and 15M samples) and a baseline without WorldPM. We fine-tune each checkpoint on Helpsteer2 and Ultrafeedback datasets, selecting models with the lowest validation loss. Our evaluation encompassed PPE metrics and downstream alignment performance assessed through BoN sampling on Qwen2.5-72B-Instruct using AlpacaEval and Arena-Hard benchmarks.

As shown in Figure 5, we observe significant improvements starting at 5M samples, with PPE-Objective showing the most consistent gains while subjective metrics exhibit more variation. This directly connects to our scaling law findings in Section 3: checkpoints with lower pre-training loss (Figure 1) consistently yield better fine-tuning performance, confirming that **larger-scale WorldPM pre-training translates into stronger downstream preference learning**.

5 Conclusion

This work demonstrates that preference modeling follows scaling laws similar to language modeling. Training on 15M StackExchange preference pairs across 1.5B to 72B models, we show that preference capabilities exhibit diverse scaling patterns, ranging from stable power-laws to capacity-gated scaling that requires sufficient model size. We also reveal an asymmetric learning mechanism in preference modeling. Furthermore, we show

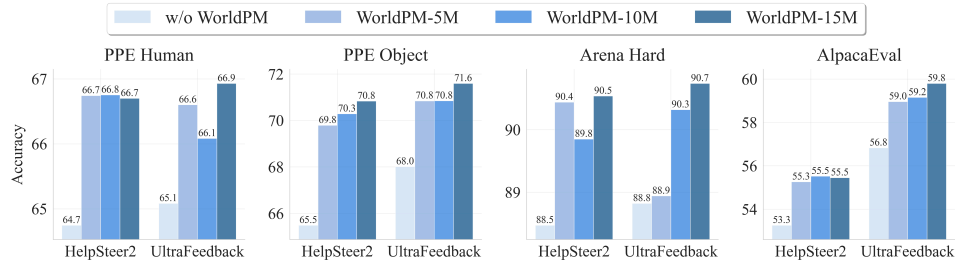


Figure 5: Comparison of PM fine-tuning performance across different WorldPM training scales. Larger WorldPM scales demonstrate enhanced fine-tuning benefits.

that measured surface confounds complicate subjective scaling evaluation, and validate WorldPM as an initialization for subsequent fine-tuning on curated or expert preferences. Future work should study confound-reduced subjective evaluation, multidimensional and intransitivity-aware preference models (Gu et al., 2021; Duan et al., 2017), and controlled scaling protocols for GenRMs.

References

- Amanda Askeell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova Dassarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, John Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, and Jared Kaplan. A general language assistant as a laboratory for alignment. *ArXiv*, abs/2112.00861, 2021. URL <https://api.semanticscholar.org/CorpusID:244799619>.
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. Program synthesis with large language models. *ArXiv*, abs/2108.07732, 2021. URL <https://api.semanticscholar.org/CorpusID:237142385>.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askeell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.2307/2334029. URL <https://doi.org/10.2307/2334029>.
- Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, et al. Rm-r1: Reward modeling as reasoning. *arXiv preprint arXiv:2505.02387*, 2025.
- Cheng-Han Chiang and Hung yi Lee. Can large language models be an alternative to human evaluations?, 2023. URL <https://arxiv.org/abs/2305.01937>.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.
- Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*, 2023.

- H. Cramér. *Mathematical Methods of Statistics*, volume 9 of *Princeton Mathematical Series*. Princeton University Press, 1946. ISBN 9780691080048. URL https://books.google.com/books?id=_db1jwECAAJ.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. ULTRAfeedback: Boosting language models with scaled AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9722–9744. PMLR, 2024. URL <https://proceedings.mlr.press/v235/cui24f.html>.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shironi Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL <https://doi.org/10.1038/s41586-025-09422-z>.
- Jiuding Duan, Jiyi Li, Yukino Baba, and Hisashi Kashima. A generalized model for multidimensional intransitivity. In *Advances in Knowledge Discovery and Data Mining: 21st Pacific-Asia Conference*, volume 10235 of *Lecture Notes in Computer Science*, pp. 840–852. Springer, 2017. doi: 10.1007/978-3-319-57529-2_65.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.
- David Easley and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010. doi: 10.1017/CBO9780511761942. URL <https://doi.org/10.1017/CBO9780511761942>.
- Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios Nikolas Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph Gonzalez, and Ion Stoica. How to evaluate reward models for rlhf. *ArXiv*, abs/2410.14872, 2024. URL <https://api.semanticscholar.org/CorpusID:273502060>.
- Xiyan Fu and Wei Liu. How reliable is multilingual LLM-as-a-judge? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11040–11053. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.587.
- Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665 – 673, 2020. URL <https://api.semanticscholar.org/CorpusID:215786368>.
- Aaron Grattafiori, Abhimanyu Dubey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL <https://arxiv.org/abs/2407.21783>.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on LLM-as-a-judge. *The Innovation*, 7(6):101253, 2026. doi: 10.1016/j.xinn.2025.101253.
- Yan Gu, Jiuding Duan, and Hisashi Kashima. An intransitivity model for matchup and pairwise comparison. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 692–698. IEEE, 2021. doi: 10.1109/ICPR48806.2021.9412462.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- Ari Holtzman, Peter West, Vered Schwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn’t always right. *ArXiv*, abs/2104.08315, 2021. URL <https://api.semanticscholar.org/CorpusID:233296182>.

- Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. *2008 Eighth IEEE International Conference on Data Mining*, pp. 263–272, 2008. URL <https://api.semanticscholar.org/CorpusID:10537313>.
- Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A Smith, Yejin Choi, and Hannaneh Hajishirzi. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. *arXiv preprint arXiv:2406.09279*, 2024.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. Scaling laws for neural language models. *ArXiv*, abs/2001.08361, 2020. URL <https://api.semanticscholar.org/CorpusID:210861095>.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.96. URL <https://aclanthology.org/2025.findings-naacl.96/>.
- Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiakai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, et al. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*, 2025a.
- Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 21558–21572, 2023.
- Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. Rm-bench: Benchmarking reward models of language models with subtlety and style. *ArXiv*, abs/2410.16184, 2024. URL <https://api.semanticscholar.org/CorpusID:273507377>.
- Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*, 2025b.
- Niklas Muennighoff, Qian Liu, Armel Zebaze, Qinkai Zheng, Binyuan Hui, Terry Yue Zhuo, Swayam Singh, Xiangru Tang, Leandro von Werra, and Shayne Longpre. Octopack: Instruction tuning code large language models. *arXiv preprint arXiv:2308.07124*, 2023.
- 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.
- Aditya Pal, F. Maxwell Harper, and Joseph A. Konstan. Exploring question selection bias to identify experts and potential experts in community question answering. *ACM Trans. Inf. Syst.*, 30:10:1–10:28, 2012. URL <https://api.semanticscholar.org/CorpusID:2855569>.
- Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung Kim, and Sanghyuk Choi. Offsetbias: Leveraging debiased data for tuning evaluators. *ArXiv*, abs/2407.06551, 2024. URL <https://api.semanticscholar.org/CorpusID:271064337>.
- 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. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL <https://arxiv.org/abs/2412.15115>.

- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *ArXiv*, abs/2311.12022, 2023. URL <https://api.semanticscholar.org/CorpusID:265295009>.
- Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46280–46302. PMLR, 2024.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.
- Hao Sun, Yunyi Shen, and Jean-François Ton. Rethinking bradley-terry models in preference-based reward modeling: Foundations, theory, and alternatives. *ArXiv*, abs/2411.04991, 2024. URL <https://api.semanticscholar.org/CorpusID:273877679>.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max W.F. Ku, Kai Wang, Alex Zhuang, Rongqi “Richard” Fan, Xiang Yue, and Wenhui Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *ArXiv*, abs/2406.01574, 2024a. URL <https://api.semanticscholar.org/CorpusID:270210486>.
- Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer2: Open-source dataset for training top-performing reward models. *ArXiv*, abs/2406.08673, 2024b. URL <https://api.semanticscholar.org/CorpusID:270440126>.
- Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. Helpsteer3-preference: Open human-annotated preference data across diverse tasks and languages. *arXiv preprint arXiv:2505.11475*, 2025.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *ArXiv*, abs/2206.07682, 2022. URL <https://api.semanticscholar.org/CorpusID:249674500>.
- Chenxi Whitehouse, Tianlu Wang, Ping Yu, Xian Li, Jason Weston, Ilia Kulikov, and Swarnadeep Saha. J1: Incentivizing thinking in llm-as-a-judge via reinforcement learning. *arXiv preprint arXiv:2505.10320*, 2025.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. doi: 10.48550/arXiv.2407.10671. URL <https://arxiv.org/abs/2407.10671>.

- Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. In *International Conference on Learning Representations (ICLR)*, 2024.
- Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *ArXiv*, abs/1611.03530, 2016. URL <https://api.semanticscholar.org/CorpusID:6212000>.
- Rui Zheng, Shihan Dou, Songyang Gao, Wei Shen, Wei-Yuan Shen, Bing Wang, Yan Liu, Senjie Jin, Qin Liu, Limao Xiong, Luyao Chen, Zhiheng Xi, Yuhao Zhou, Nuo Xu, Wen-De Lai, Minghao Zhu, Rongxiang Weng, Wen-Chun Cheng, Cheng Chang, Zhangyue Yin, Yuan Hua, Haoran Huang, Tianxiang Sun, Hang Yan, Tao Gui, Qi Zhang, Xipeng Qiu, and Xuanjing Huang. Secrets of rlhf in large language models part i: Ppo. *ArXiv*, abs/2307.04964, 2023. URL <https://api.semanticscholar.org/CorpusID:259766568>.
- Enyu Zhou, Guodong Zheng, Bing Wang, Zhiheng Xi, Shihan Dou, Rong Bao, Wei Shen, Limao Xiong, Jessica Fan, Yurong Mou, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. Rmb: Comprehensively benchmarking reward models in llm alignment. *ArXiv*, abs/2410.09893, 2024. URL <https://api.semanticscholar.org/CorpusID:273346208>.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL <https://arxiv.org/abs/2311.07911>.

Declaration of Generative AI and AI-assisted technologies

During the preparation of this work, the author(s) used Large Language Models (LLMs) to edit and polish the text to improve readability, and to generate code to assist in data visualization and figure plotting. The LLMs were not used to generate research ideas, original scientific content, or conduct evaluations. The author(s) thoroughly reviewed all generated code and edited text, and take full responsibility for the final content of this publication.

Limitation

Our dataset comprises 15M preference pairs (approximately 30G tokens) from StackExchange, modest in contrast to conventional next-token prediction pre-training, which routinely utilizes datasets of trillion-token scale. Significant opportunities exist for expanding preference pretraining data through untapped sources such as various forums and social media platforms.

All scaling experiments are conducted on a single model family (Qwen2.5), which provides a comprehensive range of model sizes from 1.5B to 72B parameters. While this enables controlled comparisons across scales, the observed scaling patterns—particularly the parameter scaling trend and the capacity-gated behavior—may be specific to this architecture and may not generalize to other model families. Validation with additional architectures is needed to establish the universality of these findings.

Our training data is predominantly English-language and technically oriented, reflecting the composition of StackExchange. This introduces potential biases: the model may favor technical or professional answer styles, and its preference judgments may not transfer well to other languages, cultural contexts, or non-technical domains such as creative writing or emotional support.

More broadly, the term “World Preference” should be understood as an aspiration rather than a precise characterization. Our current data source—primarily StackExchange—represents a specific demographic (predominantly English-speaking, technically-oriented users) with shared community norms and voting conventions. The cross-domain transferability demonstrated in Table 4 (across StackOverflow, Math SE, and other boards) may partly reflect the consistency of community norms within this ecosystem rather than truly universal human preferences. Extending WorldPM to more diverse platforms, languages, and cultural contexts is an important direction for future work to assess the generality of the preference signals captured here.

In addressing subjective evaluation biases, we control only for length and Markdown formatting. Many other influential factors, including emotional preferences, tone, level of detail, phrasing style, and cultural tendencies, remain challenging to capture, underscoring the complexity of subjective assessment. The interconnected nature of various aspects in subjective evaluation makes comprehensive and granular assessment particularly challenging, as these elements prove difficult to isolate effectively. Our style control analysis (Section 3.5) therefore provides a lower bound on the true extent of surface-feature interference in subjective evaluations.

WorldPM initialization provides weaker gains in the 800K RLHFlow setting than in smaller downstream datasets. Moreover, the 72B fine-tuned models show mixed safety results, improving some metrics and regressing on others (Appendix C.3). We report these observations without assigning an unsupported causal explanation and do not claim a consistent safety benefit.

Ethics Statement

Purpose and Intended Use

WorldPM is strictly intended for academic research purposes, specifically for: (1) investigating the scaling laws of preference modeling (reward modeling), and (2) conducting preference fine-tuning experiments to enhance fine-tuning effectiveness based on the WorldPM model.

Potential Risks and Safeguards

We acknowledge the potential risks associated with using large-scale forum data, including dual-use concerns, privacy issues, and group/cultural biases. To mitigate these risks:

1. We implement rigorous data anonymization during collection, gathering only essential fields (questions, responses, and vote counts) while removing links and personal information.
2. The model is strictly limited to non-generative purposes (reward modeling), preventing direct generation of potentially copyright-infringing or privacy-compromising content.
3. We strive to collect and balance data from diverse boards and demographics to minimize bias towards specific groups or cultures.
4. We utilize both the forum’s moderation system and response voting metrics to ensure the quality and appropriateness of the training data. Specifically, content that has been flagged or removed by the forum’s automated or human moderation processes is excluded from our dataset. Additionally, we filter out responses that have received a significant number of downvotes, as these are often indicative of content that is misleading, low-quality, inappropriate, or otherwise deemed unhelpful by the community.
5. We will restrict WorldPM to academic research purposes and applications that benefit human welfare. To mitigate potential biases, we recommend fine-tuning the model based on WorldPM before deployment.

Despite our efforts to mitigate biases by collecting data from diverse sections and user groups, we acknowledge that WorldPM may exhibit inherent biases because its StackExchange training data is mostly in English and focused on technical topics. While we have identified and mitigated certain stylistic biases (such as length preferences) through our evaluation methodology, we recognize that unidentified biases may persist. For example, the model may favor technical or professional answers over creative ones, and its focus on English means it may not perform as well on other languages.

Data Licensing and Usage

This research incorporates data from three public platforms: StackExchange, Reddit, and Quora. While Reddit and Quora data were only used for data source selection analysis in the Appendix A.1, StackExchange serves as the sole data source for model training and analysis. The model we plan to release will be trained exclusively on StackExchange data.

1. StackExchange data is publicly available through data dumps² and is licensed under the CC BY-SA 4.0 license. We strictly adhere to this license and will release our model under the same license as required.
2. Reddit provides public access to their official API³, through which we obtain their forum data. We fully adhered to Reddit’s API Terms of Service.
3. We parse Quora corpus from Common Crawl⁴. Given Quora’s Terms of Service, we will not release any Quora-related datasets or models. The Quora data used in our research has been anonymized and is limited to data source comparison purposes only.

²<https://archive.org/details/stackexchange>

³<https://www.reddit.com/dev/api>

⁴<https://commoncrawl.org/>

Reproducibility Statement

Algorithmically, WorldPM follows the same training approach as traditional BT models, with detailed training and evaluation settings provided in Appendix B. The training data for WorldPM is sourced from StackExchange, with data acquisition methods and quality enhancement through filtering mechanisms described in Section 5. Sections 3.2 and A provide comprehensive explanations of our data collection process and detailed data analysis. To ensure reproducibility, we fully disclose all details regarding training, evaluation, and data.

A Detail of Data

A.1 Training Data Selection

We collect 800K preference pairs from each forum and train them on Qwen2.5-7B to compare data source quality. As shown in Table 2, StackExchange significantly outperforms Reddit and Quora across all evaluation sets, despite being entirely out-of-domain from downstream general preference evaluations. PPE-Objective in the table is the average of five PPE objective preference evaluation sets; RMB, RewardBench, and RM-Bench report the average of their respective subsets.

Table 2: Performance comparison across forum data sources. StackExchange outperforms the other compared sources. Math SE also transfers across the evaluated domains despite its domain-specific content.

Dataset	PPE-Human	PPE-Objective	RMB	RewardBench	RM-Bench
Quora	60.2	57.7	66.1	69.0	60.9
Reddit	55.1	59.4	59.5	77.0	60.7
StackExchange	62.8	62.1	76.7	84.4	72.5
$\hookrightarrow$ Math SE	62.9	62.0	75.0	83.3	75.0

To further understand StackExchange’s out-of-domain generalization ability, we consider an extreme data source: StackExchange’s Math board, which contains only mathematics-related content. Models trained on this board perform similarly to those trained on full StackExchange data across the evaluated preference benchmarks (Table 2), indicating

cross-domain transfer within this evaluation suite.

To further assess cross-source generalization, we train models separately on each forum and evaluate cross-source prediction accuracy, as shown in Figure 6. **StackExchange shows the best generalization ability among all sources**, transferring well to Reddit, Quora, and even human-annotated HelpSteer2.

Scope of the community signal. StackExchange’s moderation, reputation-gated voting, and domain organization motivate its selection, while Tables 2–4 provide source, vote-margin, and cross-domain robustness checks. These checks support transfer beyond a single board, but they do not turn StackExchange norms into universal human preference; English-language technical-community bias remains a limitation.

A.2 Forum Data Analysis

Different dimensions of forum data generally follow long-tail distributions. We analyze all the data collected from StackExchange, as shown in Figure 7. Our analysis focuses on three key dimensions: reply length, reply upvotes, and the number of replies per post. We find that forum data, or naturally generated human data, typically exhibits strong long-tail characteristics. For instance, most forum replies are under 1K in length, receive fewer than 10 upvotes, and posts typically have fewer than 6 replies. Notably, the majority of posts have only one reply, and many replies have zero upvotes. Consequently, we need to filter out a significant portion of data where preference relationships cannot be defined, as establishing a preference pair requires at least two replies to a post with different numbers of upvotes.

Train Dataset	SE	72.7	56.2	61.8	68.4
	Reddit	64.8	60.7	58.7	59.9
	Quora	67.2	54.6	68.6	60.7
	Help	65.0	51.9	62.4	75.9
		SE	Reddit	Quora	Help
		Test Dataset			

Strong
 ↑
 Weak

Figure 6: Cross-generalization across StackExchange, Reddit, Quora, and HelpSteer2. Models trained on one source are evaluated on all others (values: test accuracy; colors: column-normalized).

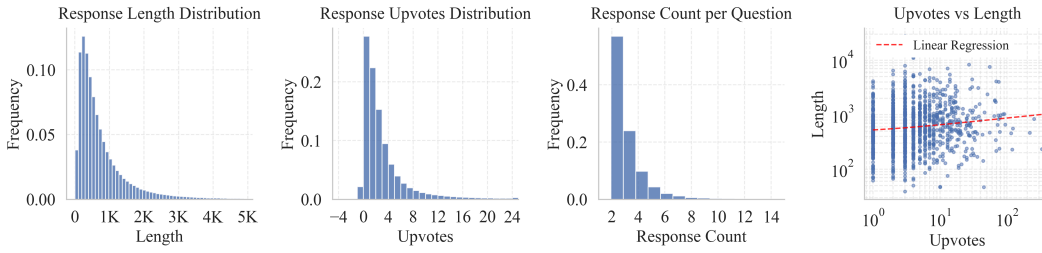


Figure 7: Distribution of reply lengths, upvotes, and the number of replies per post across all StackExchange forum data. All these characteristics demonstrate evident long-tail distributions. The last subplot indicates a weak correlation between reply length and upvotes.

It’s worth noting that there are relatively few extremely short replies in the forum data, as shown in the first graph. We hypothesize this is due to StackExchange’s relatively strict content moderation mechanisms, which filter out replies lacking substantial information.

We further analyze the relationship between reply length and upvotes, as shown in the last column of the figure. Overall, the distribution of reply lengths appears relatively uniform across different upvote counts, although there remains a positive correlation between reply length and upvotes. This correlation is natural in human-generated content, as longer replies typically indicate greater effort from the author, potentially resulting in higher quality content.

A.3 Preference Data Construction

On StackExchange, users can upvote or downvote replies, and post authors can choose to accept specific replies. We incorporate the author’s acceptance as an additional upvote. The final score for each reply is calculated by subtracting downvotes from upvotes. Using these scores, we define preference directions: when two replies have different scores, we consider them to form a preference relationship.

Score differences do not affect preference performance. One potential concern is whether replies with close scores (e.g., one reply with 1 upvote versus another with 2 upvotes) lack sufficient distinction. To address this, we bucket the score differences between reply pairs into groups: 1-2, 3-5, 6-10, and 11+ differences. As shown in Table 3, we find similar performance across different benchmarks for each bucket, with no bucket showing significant performance variations.

Similar performance across different topic domains. Another potential concern is that different boards of StackExchange, covering diverse content areas, might affect downstream generalization, especially given that our downstream evaluation spans general dialogue,

Table 3: Bucketing preference pairs by score differences to evaluate the impact of score gaps on preference data quality.

Score Diff	PPE-Human	PPE-Objective	RMB	RewardBench	RM-Bench
1-2	63.1	59.4	73.8	79.9	75.2
3-5	62.6	59.5	76.7	84.2	73.6
6-10	62.3	59.3	76.1	85.5	73.2
≥ 11	62.5	60.5	74.2	85.2	71.1

mathematics, coding, and other aspects. To address this, we divide StackExchange data into three segments: StackOverflow (computer science-related topics), Math StackExchange (mathematics-related topics), and Others. This division reflects that the majority of StackExchange content comes from the first two sections, while the remaining 170+ sections contain diverse topics with relatively few questions each.

We train models separately on each section and evaluate them across all test sets, as shown in Table 4. Our results indicate no significant performance variations across different sections. Furthermore, StackExchange provides topic tags for each question, allowing for more granular topic identification. For instance, StackOverflow uses specific tags like c++, java, and pointers. These tags also follow a long-tail distribution, with common programming questions constituting the majority. We conduct additional experiments with topic resampling (StackOverflow-TR) on StackOverflow data, deliberately oversampling from less common tags to enhance data diversity. As shown in the table, this resampling approach yields similar results.

These findings show that the learned StackExchange aggregate signal transfers across the evaluated topic domains; they do not establish that human preferences are domain-agnostic.

Table 4: Performance across StackExchange categories and topic distributions, showing cross-domain transfer within the evaluated community ecosystem.

Category	PPE-Human	PPE-Objective	RMB	RewardBench	RM-Bench
StackExchange	62.8	62.1	76.7	84.4	72.5
Other	62.7	62.0	76.2	85.7	72.3
Math StackExchange	62.9	62.0	75.0	83.3	75.0
StackOverflow	63.3	63.0	75.0	83.3	74.9
StackOverflow-TR	62.4	62.8	75.3	84.8	73.6

A.4 Aggregate Score Concentration and Constructed Comparisons

The following descriptive statistic is computed from aggregate response scores. We group pairs by the sum of the two scores and report the chosen response’s share of that sum. The available analysis record does not specify how pairs with negative scores or non-positive denominators were handled; Table 5 should therefore be read as a conditional summary of the processed pairs, not as individual-level agreement or annotator consensus.

Table 5: Aggregate score concentration in the processed preference pairs.

Combined-score bucket	Pair proportion	Mean chosen-score share
1-2	31.1%	1.000
3-5	32.3%	0.810
6-20	28.1%	0.796
21-50	5.8%	0.826
51+	2.7%	0.879

Because each retained comparison points from a lower aggregate score to a higher one, the resulting comparison graph is acyclic by construction (ties are not retained). This says

nothing about whether individual voters agree, whether their preferences are transitive, or whether aggregation hides pluralistic or non-transitive population structure.

A.5 13-gram Decontamination Check

We screen all 15M StackExchange training pairs for exact 13-gram overlap with the six domain benchmarks used to support the capacity-gated conclusion. For each automatically flagged item, we manually inspect the matching text and its context.

Table 6: Exact 13-gram screening results for the six capacity-gated domain benchmarks. “Confirmed pairs” counts benchmark-derived training pairs found by manual review, not all lexical matches.

Benchmark	Items	Flag rate (%)	Confirmed pairs
GPQA	500	0.038	0
HumanEvalPack	228	0.310	0
IFEval	500	0.015	0
MATH	529	0.394	0
MBPP-Plus	500	0.286	0
MMLU-Pro	500	0.087	0

The flagged overlaps were generic formulas, code templates, common phrases, or numeric sequences; no confirmed benchmark-derived training pair was found among them. This protocol is limited to exact lexical overlap and these six benchmarks. It does not justify a 0% contamination estimate and cannot exclude paraphrased, translated, or semantic leakage.

B Detail of Settings

B.1 Experimental Settings

We conduct world preference modeling experiments on Qwen2.5 models ranging from 1.5B to 72B parameters, with a batch size of 10K, training steps of 1536, and a learning rate of 3e-6. For smaller datasets and comparative experiments, we typically set the batch size to 2048. All experiments are performed on a cluster of 128×A100 GPUs. We use the Adam optimizer and employ learning rate warmup with a ratio of 0.1 and the weight decay coefficient of 0.1. We employ learning rate warmup with a ratio of 0.1, followed by a constant learning rate. The context length is set to 2048.

We discover that when modeling preferences, larger batch sizes lead to better performance under the same number of training steps and dataset size, as shown in Figure 8 and Figure 9. Performance continues to improve even with batch sizes up to 40K, indicating that preference modeling is indeed a challenging task that benefits from more accurate loss estimation through larger batch sizes. However, considering the limited training data available, we ultimately adopt a batch size of 10K to minimize loss noise while maintaining computational efficiency.

We additionally conduct comprehensive ablation studies on learning rates, as shown in Figure 10. Our systematic comparison of learning rates at 1e-6, 3e-6, and 1e-5 demonstrates that model performance remains stable across this range, with 3e-6 exhibiting marginally superior results. This finding suggests that preference modeling is robust to learning rate variations within reasonable bounds. However, larger learning rates require more all-compute time, which is not practical for large-scale training.

B.2 Evaluation Methods

To comprehensively evaluate WorldPM, we utilize different domain test sets from multiple RM benchmarks. Given our specific evaluation needs, we do not strictly follow the evalu-

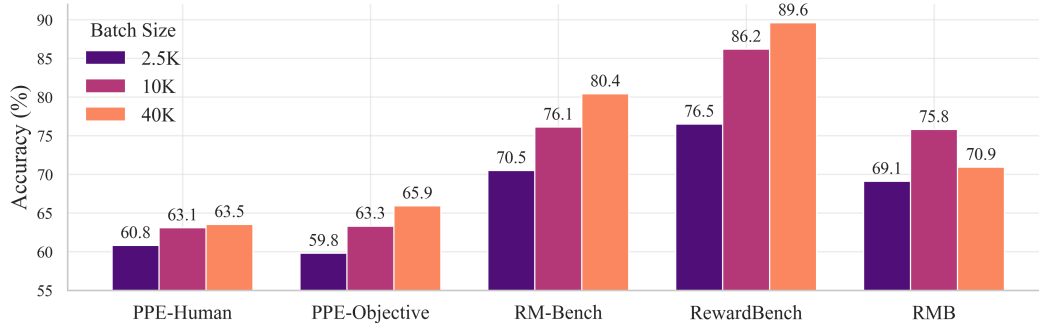


Figure 8: Comparison of different batch sizes (from 2.5K to 40K) on Qwen2.5 7B with fixed 256 steps. Larger batch sizes consistently yield better performance.

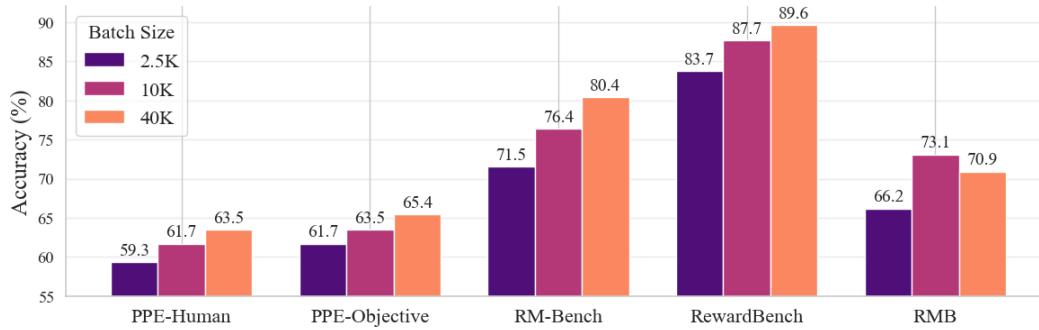


Figure 9: Comparison of different batch sizes (from 2.5K to 40K) on Qwen2.5 7B with fixed dataset size (10M). Larger batch sizes consistently yield better performance.

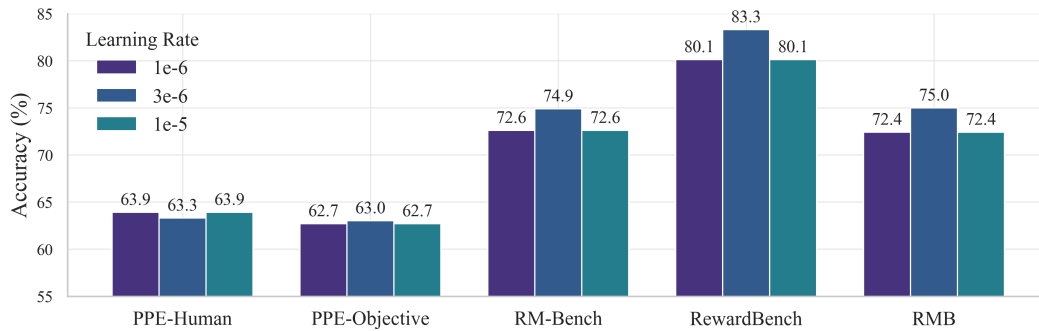


Figure 10: Through our experiments with different learning rates on Qwen2.5 7B, we observe that the model’s final performance remains relatively stable across various learning rates, with 3e-6 showing slightly superior results. This suggests that the model’s performance is robust to learning rate variations within a reasonable range.

ation protocols provided with these benchmarks. Detailed explanations can be found in Appendix B.3. The benchmarks used in this work are as follows:

- **PPE** (Frick et al., 2024): This includes evaluations of both subjective and objective parts. The data for the subjective part comes from real user annotations in the Chatbot Arena⁵. The objective part collects queries from the MMLU-Pro (Wang et al., 2024a), IFEval (Zhou et al., 2023), GPQA (Rein et al., 2023), MATH (Hendrycks et al., 2021), and MBPP-Plus (Austin et al., 2021; Liu et al., 2023) datasets, and gathers responses from state-of-the-art models. The correctness of the responses is validated against real answers, thereby forming preference pairs (with *chosen* as correct and *rejected* as incorrect).
- **RMB** (Zhou et al., 2024): This benchmark relies on GPT4 as the primary annotator, with additional human verification processes. It encompasses diverse scenarios and is fundamentally divided into two aspects: helpfulness and harmlessness.
- **RM-Bench** (Liu et al., 2024): This consists of evaluations in four domains: chat, code, math, and safety. The chat domain evaluation assesses models’ ability to identify factual errors by inserting them into responses. Prompts for code and math are sourced from the HumanEvalPack (Muennighoff et al., 2023) and MATH (Hendrycks et al., 2021) datasets, respectively, with validation against real answers. The safety section includes both pseudo-harmful and genuinely harmful questions to evaluate the model’s safety assessment capability.
- **Reward Bench** (Lambert et al., 2025): This includes evaluations in four domains: chat, chat-hard, reasoning, and safety. The chat-hard section, primarily sourced from LLMBar (Zeng et al., 2024), challenges reward models through the construction of subtly flawed responses designed to mislead evaluation.
- **Offset Bias** (Park et al., 2024): This dataset constructs high-quality but incorrect responses (rejected responses) to challenge reward models, including off-topic responses and responses containing errors.
- **HelpSteer2** (Wang et al., 2024b): This dataset, carefully annotated and filtered by trained human annotators, serves dual purposes: while it will be used as a training set in later sections, during this phase it functions as a test set for evaluating WorldPM’s subjective performance.

Task-level verifiability. Code tasks use execution-based checks, mathematics tasks use answer matching, and IFEval uses instruction-constraint checks; other benchmark components rely on benchmark-specific expert, human, or model-assisted validation. We report this categorization per benchmark rather than an item-level percentage, because we did not independently audit the verification path of every item. The StackExchange training pairs themselves have no task ground-truth labels; they are defined only by aggregate community scores.

B.3 Evaluation Settings

We evaluate our preference models using multiple RM benchmarks, including PPE, RMB, RM-Bench, and RewardBench. However, we modify some of their evaluation methodologies for our specific context.

RMB offers PairWise and Best-of-N (BoN) evaluations, where BoN measures a model’s ability to select the best reply from multiple options, and PairWise is essentially BoN with N=2. We focus on PairWise evaluation because: (1) BT loss computation inherently involves pairs of replies, making BoN evaluation

Table 7: Correlation coefficients between response length and answer quality across different models.

Model	Correlation
Gemma-2-9B-IT	+0.047
GPT-4o-Mini	−0.133
Llama-3-8B-Instruct	−0.115
Claude-3-Haiku	−0.619

⁵<https://lmarena.ai/>

computationally challenging, and (2) the original paper shows strong correlation between PairWise and BoN results.

For RM-Bench, we utilize only the model’s original responses to form preference pairs, rather than their style-modified versions, as we propose our own method for style-content separation evaluation in this paper.

For PPE, we identify a significant bias issue in the MBPP-Plus test set. This test set is constructed by sampling from four advanced models, where each model generates both correct and incorrect answers to form preference pairs. In our analysis, we observe unusually volatile metrics and conduct a detailed investigation. We define two binary variables: one indicating which answer is correct, and another indicating which answer is longer.

By measuring the correlation (Phi coefficient) between answer correctness and length, we find that Claude-3-Haiku samples show a strong negative correlation (-0.6) between these variables, where shorter answers are predominantly correct. This correlation significantly deviates from other models’ patterns, indicating a severe style bias. Consequently, we exclude Claude-3-Haiku samples from MBPP-Plus test set to ensure reliable evaluation results.

B.4 Evaluator Prompts

The evaluation prompts used in Arena Hard and Alpaca Eval are shown in Figure 11 and Figure 12, respectively.

C Additional Experimental Results

C.1 Further details on the scaling law

For each task, we fit a power-law relationship of the form $L = \alpha D^\beta$ and report the corresponding R^2 values (Figure 13). For domain preference tasks, more tasks exhibit a high goodness-of-fit as model size increases, indicating that their performance scaling can be well described by a simple power law.

We further present compute-normalized views (Figure 14), where we compare scaling trends under matched computational budgets. Overall, allocating more computation tends to yield better performance. Notably, for some tasks (e.g., GPQA, IFEval), a sufficiently large base model is required to effectively leverage additional computation; smaller models do not benefit as much from the same increase in compute. This observation is consistent with, and further supports, our findings regarding the capacity-gated scaling on objective tasks.

C.2 Scaling Trends Across All Test Sets

Figure 15 illustrates the comprehensive scaling trends of world preference across all evaluation benchmarks. In HelpSteer2’s five-dimensional assessment framework, we observe strong correlations among Helpfulness, Correctness, and Coherence dimensions, while Complexity and Verbosity demonstrate high mutual correlation. The latter two metrics show strong association with response length, as discussed in Section D.2. As elaborated in the main text, WorldPM demonstrates progressive mitigation of length preference, manifesting as increasing loss on these metrics.

Safety evaluation, omitted from the main text due to space constraints, encompasses RMB-Harmlessness, RewardBench-Safety, and RM-Bench-Safety metrics. RewardBench-Safety exhibits distinct scaling patterns across model sizes: the 1.5B model shows continuous decline, while 7B and 72B models stabilize post-decline, achieving accuracy exceeding 90%, indicating performance saturation. RMB-Harmlessness presents heterogeneous patterns: declining for 1.5B, saturating for 7B, and ascending for 72B. These diverse patterns can be partially interpreted through RM-Bench-Safety’s subdomain analysis: pseudo-harmful queries versus genuinely harmful queries (RM-Bench-Safety-Response), as RMB-Harmlessness also

Arena Hard's Prompt

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user prompt displayed below. You will be given assistant A's answer and assistant B's answer. Your job is to evaluate which assistant's answer is better.

Begin your evaluation by generating your own answer to the prompt. You must provide your answers before judging any answers.

When evaluating the assistants' answers, compare both assistants' answers with your answer. You must identify and correct any mistakes or inaccurate information.

Then consider if the assistant's answers are helpful, relevant, and concise. Helpful means the answer correctly responds to the prompt or follows the instructions. Note when user prompt has any ambiguity or more than one interpretation, it is more helpful and appropriate to ask for clarifications or more information from the user than providing an answer based on assumptions. Relevant means all parts of the response closely connect or are appropriate to what is being asked. Concise means the response is clear and not verbose or excessive.

Then consider the creativity and novelty of the assistant's answers when needed. Finally, identify any missing important information in the assistants' answers that would be beneficial to include when responding to the user prompt.

After providing your explanation, you must output only one of the following choices as your final verdict with a label:

1. Assistant A is significantly better: `[[A>>B]]`
2. Assistant A is slightly better: `[[A>B]]`
3. Tie, relatively the same: `[[A=B]]`
4. Assistant B is slightly better: `[[B>A]]`
5. Assistant B is significantly better: `[[B>>A]]`

Example output: "My final verdict is tie: `[[A=B]]`".

Figure 11: Evaluation prompts used in Arena Hard.

incorporates both aspects. Our analysis reveals decreasing and saturating loss patterns for genuine harmful content detection, while pseudo-harmful query response shows ascending loss in later training stages. These metrics move in different directions; the present experiments do not identify the cause of that divergence.

C.3 Preference Fine-tuning Evaluation

Tables 8 and 9 present uncontrolled preference fine-tuning results for 7B and 72B models with and without WorldPM initialization, while Tables 10 and 11 show the style-controlled versions. The results align with findings in main text.

Notably, we include safety evaluations (RMB-Harmless, RewardBench-Safety, and RM-Bench-Safety), where the 72B model shows mixed performance without consistent improvement. We do not assign a causal explanation to this pattern. Detailed results are in Appendix C.2.

Alpaca Eval's Prompt

```

<|im_start|>system
You are a highly efficient assistant, who evaluates and selects the
  best large language model (LLMs) based on the quality of their
  responses to a given instruction. This process will be used to
  create a leaderboard reflecting the most accurate and human-
  preferred answers.
<|im_end|>
<|im_start|>user
I require a leaderboard for various large language models. I'll
  provide you with prompts given to these models and their
  corresponding outputs. Your task is to assess these responses,
  and select the model that produces the best output from a human
  perspective.

## Instruction

{
  "instruction": "{instruction}",
}

## Model Outputs

Here are the unordered outputs from the models. Each output is
  associated with a specific model, identified by a unique model
  identifier.
{
  {
    "model_identifier": "m",
    "output": "{output_1}"
  },
  {
    "model_identifier": "M",
    "output": "{output_2}"
  }
}

## Task

Evaluate the models based on the quality and relevance of their
  outputs, and select the model that generated the best output.
  Answer by first providing a concise explanation and then end
  your answer by providing the model identifier of the best output
  . We will use the last character of your output `output[-1]` as
  the name of the best model, so make sure you finish with the
  token of the model identifiers and nothing else: `m` or `M` (no
  quotes, no dots, no backticks, no new lines,...). For example:

### Concise explanation
... some text...
### Which is best, m or M?
M

Now is your turn.

## Your answer: "Concise explanation" followed by "Which is best, m
  or M?"
<|im_end|>

```

Figure 12: Evaluation prompts used in Alpaca Eval.

Table 8: Effect of WorldPM initialization on 7B model’s PM fine-tuning performance across different evaluation categories (without style control). Abbreviations: RB - RewardBench, RM-B - RM-Bench, Safety-R/F - Safety-Response/Refuse.

Metrics	HelpSteer2		UltraFeedback		RLHFlow	
	w/o WorldPM	w/ WorldPM	w/o WorldPM	w/ WorldPM	w/o WorldPM	w/ WorldPM
Subjective Evaluation						
PPE-Human	62.05	63.80 (↑2.81%)	60.88	61.64 (↑1.25%)	65.54	66.03 (↑0.75%)
RMB-Helpful	72.49	75.84 (↑4.62%)	75.62	75.70 (↑0.11%)	77.64	78.33 (↑0.88%)
RB-Chat	97.49	97.21 (↓0.29%)	97.49	97.21 (↓0.29%)	98.04	97.77 (↓0.28%)
Domain Preference Evaluation						
PPE-MBPP	67.07	71.31 (↑6.33%)	66.60	72.24 (↑8.46%)	71.97	74.44 (↑3.43%)
PPE-GPQA	55.41	56.56 (↑2.08%)	55.62	55.18 (↓0.81%)	56.76	56.70 (↓0.10%)
PPE-IFEval	57.50	56.17 (↓2.31%)	56.21	56.64 (↑0.76%)	61.87	62.54 (↑1.07%)
PPE-MATH	70.80	71.74 (↑1.32%)	71.64	73.77 (↑2.97%)	78.79	79.94 (↑1.46%)
PPE-MMLU	63.20	66.21 (↑4.76%)	65.51	66.17 (↑1.01%)	67.56	70.20 (↑3.90%)
RM-B-Code	55.92	57.68 (↑3.14%)	57.02	62.06 (↑8.85%)	61.40	67.76 (↑10.36%)
RM-B-Math	69.09	74.10 (↑7.25%)	70.23	73.44 (↑4.58%)	75.99	79.58 (↑4.73%)
RB-Reason	81.62	96.12 (↑17.77%)	83.19	91.86 (↑10.42%)	94.13	96.16 (↑2.15%)
Common-sense Preference Evaluation						
RB-Chat-Hard	62.50	74.89 (↑19.82%)	61.51	67.32 (↑9.45%)	66.01	69.96 (↑5.98%)
RM-B-Chat	51.94	65.12 (↑25.37%)	57.36	66.67 (↑16.22%)	56.98	66.67 (↑17.01%)
OffsetBias	63.76	82.40 (↑29.23%)	72.98	78.17 (↑7.11%)	73.54	80.50 (↑9.47%)
Safety Evaluation						
RMB-Harmless	68.41	73.58 (↑7.56%)	69.22	70.23 (↑1.46%)	66.78	66.84 (↑0.08%)
RB-Safety	82.84	90.54 (↑9.30%)	89.05	87.84 (↓1.37%)	85.95	86.35 (↑0.47%)
RM-B-Safety-R	87.90	92.99 (↑5.80%)	95.54	94.27 (↓1.33%)	91.08	93.63 (↑2.80%)
RM-B-Safety-F	98.59	97.54 (↓1.07%)	96.48	96.48 (0.00%)	95.77	95.42 (↓0.37%)

Table 9: Effect of WorldPM initialization on 72B model’s PM fine-tuning performance across different evaluation categories (without style control). Abbreviations: RB - RewardBench, RM-B - RM-Bench, Safety-R/F - Safety-Response/Refuse.

Metrics	HelpSteer2		UltraFeedback		RLHFlow	
	w/o WorldPM	w/ WorldPM	w/o WorldPM	w/ WorldPM	w/o WorldPM	w/ WorldPM
Subjective Evaluation						
PPE-Human	62.22	65.53 (↑5.32%)	62.09	64.51 (↑3.89%)	67.63	68.29 (↑0.98%)
RMB-Helpful	76.40	77.45 (↑1.38%)	75.60	79.10 (↑4.64%)	78.10	80.08 (↑2.53%)
RB-Chat	97.91	98.88 (↑1.00%)	97.21	96.09 (↓1.15%)	98.32	98.04 (↓0.28%)
Domain Preference Evaluation						
PPE-MBPP	70.19	77.45 (↑10.34%)	74.32	78.65 (↑5.82%)	81.00	81.85 (↑1.05%)
PPE-GPQA	58.03	59.28 (↑2.15%)	59.06	60.76 (↑2.88%)	62.19	63.44 (↑2.01%)
PPE-IFEval	58.96	63.50 (↑7.68%)	60.16	63.83 (↑6.10%)	63.32	65.04 (↑2.71%)
PPE-MATH	69.73	76.64 (↑9.92%)	72.58	77.89 (↑7.32%)	79.10	82.52 (↑4.32%)
PPE-MMLU	69.98	74.69 (↑6.73%)	72.07	74.92 (↑3.96%)	78.52	79.28 (↑0.97%)
RM-B-Code	61.18	69.30 (↑13.26%)	65.79	68.64 (↑4.33%)	71.93	72.81 (↑1.22%)
RM-B-Math	70.04	76.75 (↑9.58%)	71.55	77.32 (↑8.06%)	87.52	85.63 (↓2.16%)
RB-Reason	96.61	98.22 (↑1.66%)	95.70	96.44 (↑0.77%)	96.33	97.38 (↑1.09%)
Common-sense Preference Evaluation						
RB-Chat-Hard	69.41	80.26 (↑15.64%)	75.66	77.63 (↑2.61%)	74.12	75.33 (↑1.63%)
RM-B-Chat	77.52	75.97 (↓2.00%)	70.93	70.54 (↓0.55%)	77.13	75.58 (↓1.01%)
OffsetBias	76.29	85.79 (↑12.46%)	84.47	85.86 (↑1.64%)	83.18	81.90 (↓1.54%)
Safety Evaluation						
RMB-Harmless	68.96	70.10 (↑1.66%)	68.58	67.93 (↓0.95%)	63.70	62.65 (↓1.64%)
RB-Safety	90.14	91.35 (↑1.35%)	89.12	90.61 (↑1.67%)	84.19	83.24 (↓1.12%)
RM-B-Safety-R	94.59	94.27 (↓0.34%)	97.45	98.09 (↑0.65%)	94.90	93.63 (↓1.34%)
RM-B-Safety-F	97.89	97.18 (↓0.72%)	97.01	96.48 (↓0.54%)	88.73	89.08 (↑0.40%)

Table 10: Effect of WorldPM initialization on 7B model’s PM fine-tuning performance across different evaluation categories. Abbreviations: RB - RewardBench, RM-B - RM-Bench, Safety-R/F - Safety-Response/Refuse.

Metrics	Helpsteer2		UltraFeedback		RLHFlow	
	w/o WorldPM	w/ WorldPM	w/o WorldPM	w/ WorldPM	w/o WorldPM	w/ WorldPM
Subjective Evaluation						
PPE-Human	63.32	64.93 (↑2.54%)	63.22	63.86 (↑1.01%)	66.18	66.78 (↑0.91%)
RMB-Helpful	74.33	77.02 (↑3.61%)	78.37	78.82 (↑0.58%)	78.66	79.27 (↑0.78%)
RB-Chat	97.21	97.07 (↓0.14%)	97.49	97.21 (↓0.29%)	98.04	97.77 (↓0.28%)
Domain Preference Evaluation						
PPE-MBPP	72.24	73.55 (↑1.82%)	69.77	73.55 (↑5.42%)	75.06	76.83 (↑2.37%)
PPE-GPQA	56.27	56.70 (↑0.76%)	56.19	55.43 (↓1.36%)	57.34	57.60 (↑0.44%)
PPE-IFEval	58.81	57.34 (↓2.49%)	57.42	57.03 (↓0.68%)	62.09	62.38 (↑0.47%)
PPE-MATH	71.04	71.64 (↑0.85%)	71.76	74.36 (↑3.62%)	78.98	80.16 (↑1.48%)
PPE-MMLU	63.55	66.52 (↑4.67%)	65.64	66.66 (↑1.55%)	67.97	70.27 (↑3.39%)
RM-B-Code	57.68	60.96 (↑5.70%)	59.21	62.94 (↑6.30%)	65.35	67.76 (↑3.69%)
RM-B-Math	70.70	75.99 (↑7.49%)	72.12	75.71 (↑4.98%)	76.56	80.43 (↑5.06%)
RB-Reason	88.02	96.44 (↑9.57%)	92.35	95.39 (↑3.29%)	95.91	96.82 (↑0.95%)
Common-sense Preference Evaluation						
RB-Chat-Hard	81.25	83.99 (↑3.37%)	77.52	79.82 (↑2.97%)	81.47	82.57 (↑1.35%)
RM-B-Chat	78.29	82.17 (↑4.95%)	80.23	81.01 (↑0.97%)	82.17	82.56 (↑0.47%)
OffsetBias	78.87	86.60 (↑9.80%)	81.68	82.94 (↑1.55%)	83.43	85.69 (↑2.71%)
Safety Evaluation						
RMB-Harmless	72.73	73.99 (↑1.72%)	69.74	70.92 (↑1.69%)	62.86	63.79 (↑1.48%)
RB-Safety	90.68	94.53 (↑4.25%)	91.76	89.46 (↓2.50%)	90.88	92.03 (↑1.26%)
RM-B-Safety-R	90.45	93.95 (↑3.87%)	96.50	96.82 (↑0.33%)	92.99	96.18 (↑3.42%)
RM-B-Safety-F	99.65	99.30 (↓0.35%)	98.94	98.94 (0.00%)	98.59	99.12 (↑0.54%)
Alignment Evaluation						
Arena Hard	65.88	70.27 (↑6.66%)	67.93	72.12 (↑6.17%)	72.41	74.62 (↑3.05%)
Alpaca Eval	43.73	46.29 (↑5.84%)	46.16	46.35 (↑0.40%)	53.24	50.97 (↓4.27%)
Arena Hard-SC	65.07	70.22 (↑7.91%)	66.19	70.36 (↑6.30%)	72.77	74.29 (↑2.09%)
Alpaca Eval-LC	39.98	42.14 (↑5.41%)	44.46	46.13 (↑3.76%)	46.24	46.32 (↑0.17%)

Table 11: Effect of WorldPM initialization on 72B model’s PM fine-tuning performance across different evaluation categories. Abbreviations: RB - RewardBench, RM-B - RM-Bench, Safety-R/F - Safety-Response/Refuse.

Metrics	Helpsteer2		UltraFeedback		RLHFlow	
	w/o WorldPM	w/ WorldPM	w/o WorldPM	w/ WorldPM	w/o WorldPM	w/ WorldPM
Subjective Evaluation						
PPE-Human	64.76	66.68 (↑2.97%)	65.12	66.94 (↑2.80%)	68.14	68.93 (↑1.16%)
RMB-Helpful	77.76	78.53 (↑1.00%)	79.80	81.73 (↑2.41%)	79.26	80.89 (↑2.05%)
RB-Chat	98.04	98.88 (↑0.85%)	97.63	96.37 (↓1.29%)	98.46	97.91 (↓0.57%)
Domain Preference Evaluation						
PPE-MBPP	70.50	78.76 (↑11.72%)	75.02	79.46 (↑5.92%)	82.28	82.39 (↑0.14%)
PPE-GPQA	57.66	59.63 (↑3.42%)	59.22	61.19 (↑3.33%)	62.52	63.75 (↑1.97%)
PPE-IFEval	59.24	63.63 (↑7.42%)	60.57	64.34 (↑6.22%)	64.26	65.39 (↑1.76%)
PPE-MATH	69.94	77.17 (↑10.33%)	72.77	78.16 (↑7.41%)	79.36	83.01 (↑4.60%)
PPE-MMLU	70.20	74.96 (↑6.79%)	72.42	75.08 (↑3.67%)	79.04	79.30 (↑0.32%)
RM-B-Code	60.75	69.30 (↑14.08%)	67.32	70.18 (↑4.23%)	72.15	73.46 (↑1.82%)
RM-B-Math	72.78	77.69 (↑6.75%)	73.72	77.79 (↑5.51%)	87.81	86.20 (↓1.83%)
RB-Reason	96.65	98.25 (↑1.66%)	96.30	97.48 (↑1.23%)	97.52	98.15 (↑0.64%)
Common-sense Preference Evaluation						
RB-Chat-Hard	84.87	87.28 (↑2.84%)	84.54	83.00 (↓1.82%)	84.21	84.54 (↑0.39%)
RM-B-Chat	81.01	82.17 (↑1.44%)	80.23	79.46 (↓0.97%)	85.66	84.50 (↓1.36%)
OffsetBias	83.14	89.76 (↑7.96%)	87.29	88.05 (↑0.88%)	87.89	87.47 (↓0.47%)
Safety Evaluation						
RMB-Harmless	69.42	70.34 (↑1.33%)	69.57	68.43 (↓1.64%)	60.47	59.96 (↓0.84%)
RB-Safety	92.03	93.51 (↑1.62%)	93.45	93.51 (↑0.07%)	92.09	92.03 (↓0.07%)
RM-B-Safety-R	95.54	97.13 (↑1.67%)	98.41	98.41 (0.00%)	98.41	97.13 (↓1.29%)
RM-B-Safety-F	99.65	99.12 (↓0.53%)	98.59	98.24 (↓0.36%)	96.83	97.18 (↑0.36%)
Alignment Evaluation						
Arena Hard	88.52	89.72 (↑1.36%)	89.16	90.37 (↑1.36%)	90.41	90.82 (↑0.45%)
Alpaca Eval	60.34	63.39 (↑5.07%)	59.95	63.73 (↑6.31%)	65.51	66.26 (↑1.14%)
Arena Hard-SC	88.47	90.54 (↑2.34%)	88.82	90.74 (↑2.16%)	91.06	91.76 (↑0.77%)
Alpaca Eval-LC	53.26	55.45 (↑4.12%)	56.83	59.80 (↑5.23%)	56.95	55.58 (↓2.40%)

Task	1.5B: $L = \alpha D^\beta$ (R^2)	7B: $L = \alpha D^\beta$ (R^2)	32B: $L = \alpha D^\beta$ (R^2)	72B: $L = \alpha D^\beta$ (R^2)
Factual Error Detect	$L = -0.273D^{-0.129}$ (0.998)	$L = -0.404D^{-0.148}$ (0.991)	$L = -0.515D^{-0.106}$ (0.983)	$L = -0.600D^{-0.117}$ (0.974)
OffsetBias	$L = -0.308D^{-0.129}$ (0.991)	$L = -0.701D^{-0.093}$ (0.975)	$L = -0.664D^{-0.132}$ (0.986)	$L = -0.773D^{-0.091}$ (0.902)
LLMBar	$L = -0.239D^{-0.061}$ (0.882)	$L = -0.417D^{-0.086}$ (0.944)	$L = -0.449D^{-0.073}$ (0.919)	$L = -0.548D^{-0.056}$ (0.842)
MBPP-Plus	$L = -0.373D^{-0.005}$ (0.673)	$L = -0.491D^{-0.022}$ (0.711)	$L = -0.496D^{-0.040}$ (0.911)	$L = -0.507D^{-0.021}$ (0.903)
MMLU-Pro	$L = -0.371D^{-0.001}$ (0.027)	$L = -0.449D^{-0.017}$ (0.715)	$L = -0.466D^{-0.038}$ (0.957)	$L = -0.513D^{-0.033}$ (0.897)
IFEval	$L = -0.349D^{-0.001}$ (0.113)	$L = -0.367D^{-0.003}$ (0.633)	$L = -0.369D^{-0.004}$ (0.891)	$L = -0.384D^{-0.011}$ (0.957)
HumanEvalPack	$L = -0.362D^{-0.005}$ (0.493)	$L = -0.390D^{0.003}$ (0.310)	$L = -0.400D^{-0.017}$ (0.968)	$L = -0.419D^{-0.019}$ (0.959)
GPQA	$L = -0.372D^{0.000}$ (0.014)	$L = -0.370D^{0.004}$ (0.706)	$L = -0.380D^{-0.002}$ (0.208)	$L = -0.384D^{-0.007}$ (0.962)
MATH	$L = -0.365D^{-0.005}$ (0.389)	$L = -0.424D^{-0.041}$ (0.923)	$L = -0.492D^{-0.056}$ (0.966)	$L = -0.469D^{-0.033}$ (0.901)
Chatbot Arena	$L = -0.387D^{0.009}$ (0.803)	$L = -0.435D^{0.002}$ (0.068)	$L = -0.425D^{-0.003}$ (0.462)	$L = -0.441D^{0.004}$ (0.400)
HelpSteer2	$L = -0.388D^{0.002}$ (0.153)	$L = -0.493D^{-0.011}$ (0.453)	$L = -0.475D^{-0.013}$ (0.748)	$L = -0.529D^{-0.005}$ (0.224)
GPT4	$L = -0.464D^{0.027}$ (0.709)	$L = -0.638D^{0.009}$ (0.066)	$L = -0.586D^{-0.004}$ (0.149)	$L = -0.594D^{0.002}$ (0.018)

Figure 13: Power-law fits of the form $L = \alpha D^\beta$ for each task, along with the corresponding R^2 values.

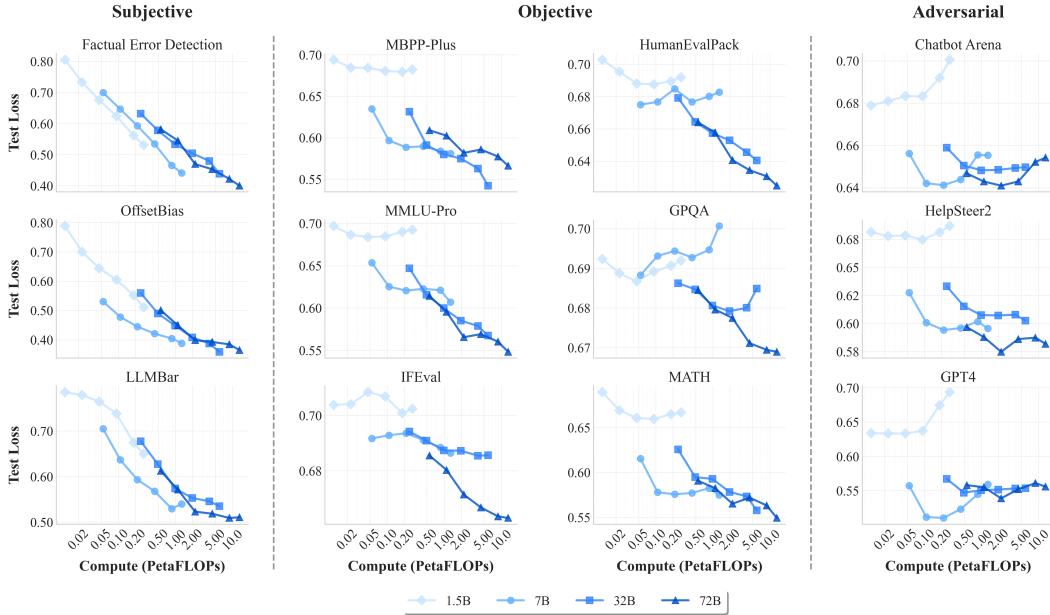


Figure 14: Compute-normalized scaling behavior across tasks. We compare performance trends under matched computational budgets. In general, increased computation leads to better performance, and certain tasks require larger base models in order to fully benefit from higher compute.

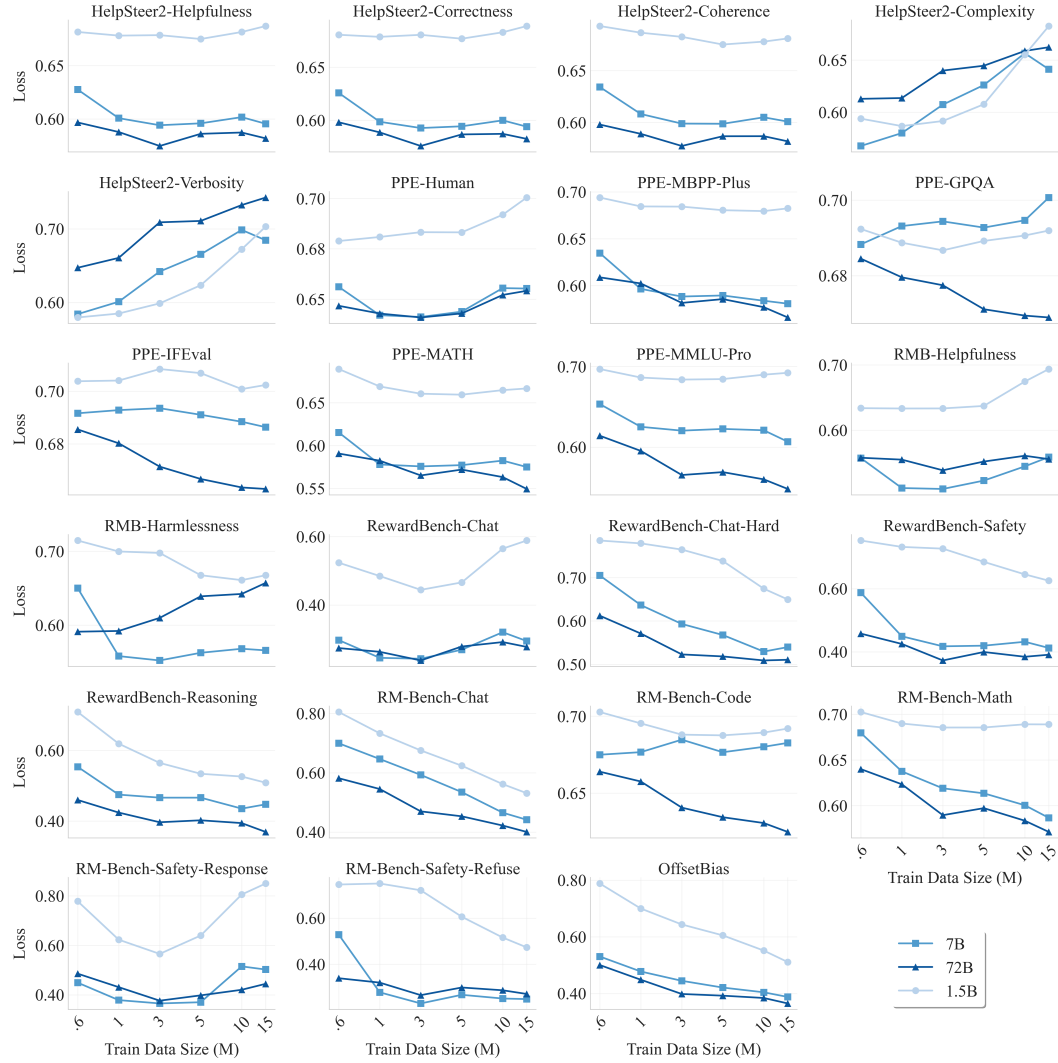


Figure 15: Loss trajectories across evaluation benchmarks as a function of training scale and model size, plotted on $\log_2$ scale.

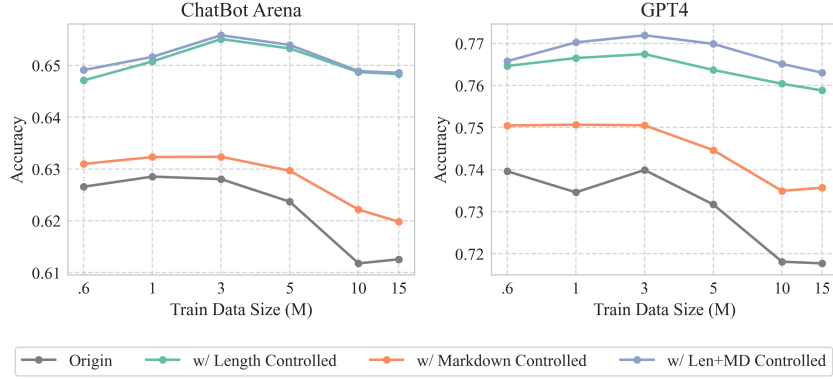


Figure 16: Impact of controlling different style factors on subjective evaluation performance.

D Details of Style Control

D.1 Ablation Study on Style Control Factors

We control two style features: length and markdown formatting. As shown in Figure 16, we compare the impact on subjective evaluation performance when controlling length alone, markdown formatting alone, and both factors simultaneously. Our findings indicate that markdown formatting has less influence than length, further confirming that length is the primary style factor. Moreover, modeling both factors together enables more effective style control.

D.2 Effect of Style Control Across Test Set Evaluation

Figure 18 and Figure 17 demonstrates the impact of style control on evaluation results (test accuracy and test loss) across all test sets. We observe several key findings:

- In HelpSteer2, the dimensions of Helpfulness, Correctness, and Coherence maintain substantial consistency. Complexity and Verbosity exhibit high correlation with length characteristics. Upon implementing length control, these two metrics show significant changes while maintaining stable trends.
- Some domain preference and common-sense preference evaluation sets demonstrate notable shifts in performance (e.g., PPE-MBPP-Plus, PPE-GPQA, and RM-Bench-Chat). However, the performance gap between style-controlled and uncontrolled versions either remains constant or narrows with increased training scale, contrasting with the widening gap in subjective domains. This suggests that modeling world preference gradually overcomes length bias, converging toward more accurate evaluation results.
- In the safety domain, RM-Bench-Safety-Refuse demonstrates exceptionally high accuracy even without style control, indicating WorldPM’s inherent capability for safety discrimination. Other safety-related benchmarks show declining accuracy in later stages, potentially related to increased defense against pseudo-harmful queries.

E The Use of Large Language Models (LLMs)

The use of LLMs in this paper is limited to polishing content and correcting errors during the writing phase. All content generated by LLMs has been thoroughly reviewed and verified.

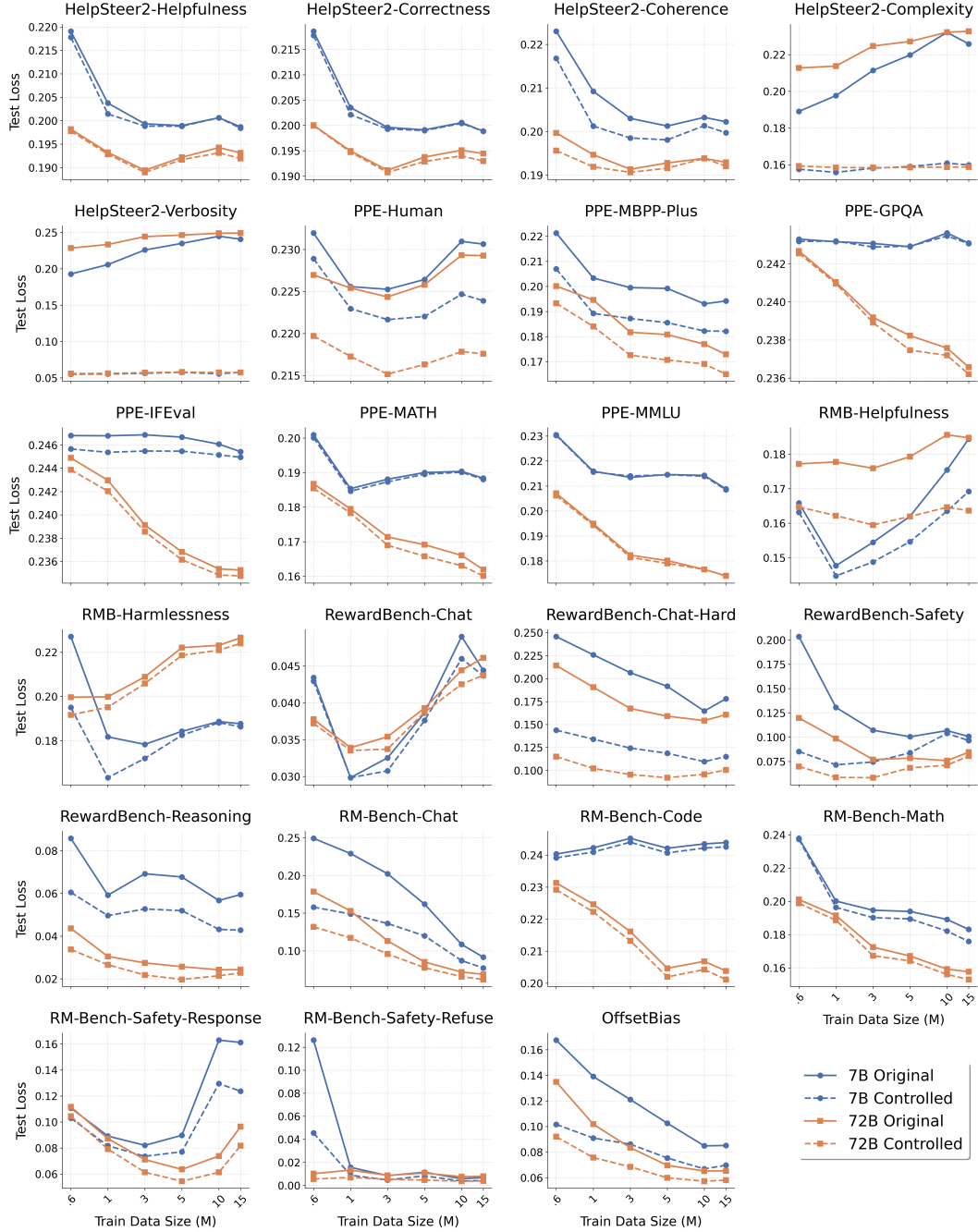


Figure 17: Effect of style control on test loss across all test set evaluation.

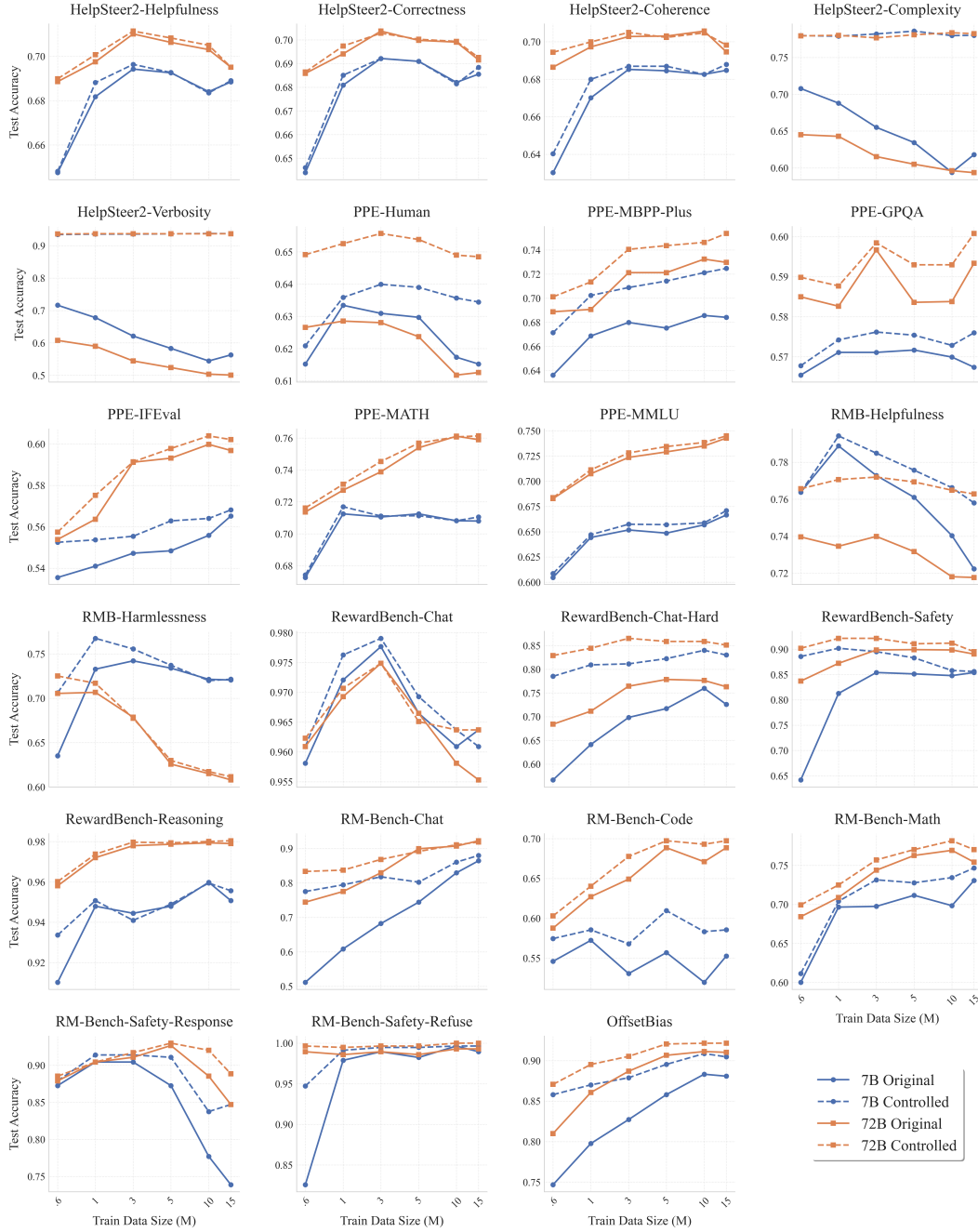


Figure 18: Effect of style control on performance across all test set evaluation.

F Case Study

We present several examples from PPE-Human, including Figures 19, 20, and 21. In Figures 19 and 20, general humans chose incorrect answers, while WorldPM correctly identified the errors. Figure 21 demonstrates that data from general humans can be low-quality and uninformative, thus interfering with evaluation.

We also present several examples of training data used for modeling world preference, including Figures 22, 23, 24, and 25. These data often contain rich information, enabling the model to learn complex representations.

Query
Prove or disprove: for all integers a, b, c, d , if $a \mid b$ and $c \mid d$, then $(a + c) \mid (b + d)$.
Chosen Response (World Preference Model Score: -1.84) This statement is true. This is because absolute values of sums of integers are equal to the sums of their absolute values. In this case, since $a \mid b$ (a divides b) and $c \mid d$ (c divides d), then certainly b and d are both divisible by a and c, respectively. So, setting $e = a$ and $f = c$, we have that both $b = q \cdot e$ for some integer q and $d = r \cdot f$ for some integer r. Then, $b + d = (q + r) \cdot e + f$ is divisible by both e and f, which are equal to a and c. So $(b + d)$ is divisible by the sum $a + c$. Thus, $(a + c) \mid (b + d)$.
Rejected Response (World Preference Model Score: 0.76) To prove or disprove the statement "for all integers a, b, c, d if $a \mid b$ and $c \mid d$, then $a + c \mid b + d$", we can approach it by first understanding what the statement means and then testing it with specific examples. Definitions: - $a \mid b$ means that there exists an integer k such that $b = ak$. - $c \mid d$ means that there exists an integer m such that $d = cm$. Claim: - $a + c \mid b + d$ means there exists an integer n such that $b + d = (a + c)n$. We need to see if from $b = ak$ and $d = cm$, it follows that $b + d = (a + c)n$ for some integer n. Counterexample: Let's consider specific numbers to see if this statement holds. - Choose $a = 2, b = 4, c = 3, d = 6$. - Clearly, $2 \mid 4$ because $4 = 2 \times 2$. - Clearly, $3 \mid 6$ because $6 = 3 \times 2$. - Now, $a + c = 2 + 3 = 5$ and $b + d = 4 + 6 = 10$. We need to check if $5 \mid 10$. Since $10 = 5 \times 2$, in this case, 5 does actually divide 10. However, we need to check more examples to see if there are any that fail. Another Example: - Choose $a = 2, b = 4, c = 3, d = 9$. - Again, $2 \mid 4$ since $4 = 2 \times 2$. - $3 \mid 9$ since $9 = 3 \times 3$. - Calculate $a + c = 2 + 3 = 5$ and $b + d = 4 + 9 = 13$. Now, we need to check if $5 \mid 13$. Since 13 is not divisible by 5 (there is no integer n such that $13 = 5n$), this proves that 5 does not divide 13. Conclusion: The counterexample where $a = 2, b = 4, c = 3, d = 9$ disproves the statement. Even though a divides b and c divides d, $a + c$ does not necessarily divide $b + d$. Thus, the statement "for all integers a, b, c, d if $a \mid b$ and $c \mid d$, then $a + c \mid b + d$" is false.

Figure 19: Example from the PPE-Human test set where the chosen response's proof is wrong while the rejected response's proof is correct, and the world preference model identify this error.

Query

userscript to copy the text of the hovered link when the key combination alt+c is pressed. just present the code and make it concise.

Chosen Response (World Preference Model Score: -0.66)

```
// @require      https://code.jquery.com/jquery-3.6.0.min.js

$(document).keydown(function(event){
  if(event.altKey && event.keyCode == 67){
    var link = $("hoveredElement").text();
    clipboard.writeText(link);
  }
});
```

Rejected Response (World Preference Model Score: 1.42)

```
// ==UserScript==
// @name         Copy Hovered Link Text
// @namespace    http://tampermonkey.net/
// @version      1.0
// @description   Copy the text of the hovered link when Alt+C is
//               pressed
// @author       You
// @match        *://*/
// @grant        none
// ==/UserScript==

(function() {
  'use strict';
  let hoveredLinkText = '';
  document.addEventListener('mouseover', function(e) {
    if (e.target.tagName.toLowerCase() === 'a') {
      hoveredLinkText = e.target.innerText || e.target.
        textContent;
    }
  });
  document.addEventListener('keydown', function(e) {
    if (e.altKey && e.key === 'c') {
      if (hoveredLinkText) {
        navigator.clipboard.writeText(hoveredLinkText).then
          (() => {
            console.log('Text copied to clipboard: ',
              hoveredLinkText);
          }).catch(err => {
            console.error('Failed to copy text: ', err);
          });
      }
    }
  });
})();
```

Figure 20: Example from the PPE-Human test set, where Chosen uses an invalid selector `$("hoveredElement")`, while Rejected provides a correct solution.

Query	
r u gpt 4o? answer is only yes or no	
Chosen Response	Rejected Response
Yes	No.

Figure 21: Example from the PPE-Human test set, where the relative quality cannot be determined.

Query (From freelancing.stackexchange)

Title: Where can a front end web developer start freelancing?

Let's say you're beginning web development, and you want to start earning money. Where do you get started?

I know a portfolio would be a start, but I don't know where to apply just as a front-end developer for jobs.

I don't know and don't want to know PHP, Ruby, Rails, and other stuff. I just want to do front end development part time, because I find it interesting.

By the way, if it helps, I live in NYC.

Chosen Response (Upvotes: 11)

The challenge that you're going to run into with Freelance web development is that, in my experience, many Web developers have pretty broad understanding of the entire development stack from databases to server-side coding to front-end HTML, CSS, and JavaScript.

Thus, if you're just starting out and wish to avoid learning other parts of the stack, you may run into trouble finding work. In my experience, when a Freelancer needs to outsource something, it's because he or she needs a specialist in that area.

Using myself as an example, I lack knowledge in design, so I'm less likely to find work building websites since most people who build websites can also manage to do the little bit of scripting necessary to make the site functional. However, I specialize in JavaScript. If you have a Web application and are stuck somewhere or are doing something complicated, that's where I come in.

If you're going to specialize in nothing but the frontend, you're going to need to specialize in it, which means you're going to need to be really good, write really clean, W3C-valid HTML code that works in every browser, and have a deep understanding of front-end User Experience.

Even then, you need to understand at least a little server-side code. The best front-end coders I've ever worked with still knew a little PHP so that we could understand how we were going to integrate things. They weren't experts in it, but they weren't ignorant to it either; it's naive to think that you can get away with never ever knowing how the server-side works.

With that said, you have to start somewhere, and the best way for you to do that is to look for small jobs that require just minor changes to existing front-end code. This will help build your skill-set in that area while hopefully giving you some income.

Rejected Response (Upvotes: 8)

If you're doing front end development, you will need a portfolio before you get any big jobs, end of story. So how do you get a portfolio without the jobs?

What I did before was make sample pages for friends of the family, and a couple volunteer organisations I belonged to. It was all under the impression that they only paid if they used my designs, but accept that it's just a start. I could at least get some styles that I could do. That portfolio should be online though on your personal website, so potential customers can see your skill, design, and technique.

Again, just make designs, no matter who for, and make sure they show your best talents

Figure 22: Example 1 drawn from StackExchange training dataset.

Query (From physics.stackexchange)

Title: Could a non-rotating planet, in a binary situation, have gravity?
Is it possible for a singular planet to rotate around a single sun in the same way that our moon rotates around the Earth? If so, would it be possible for that planet, if as large or larger than Earth, to have gravity?

Chosen Response (Upvotes: 4)

Yes. Literally everything that has mass (and even some things that don't) will "have gravity". It is certainly possible for a single planet to orbit a single star (as opposed to our planet, for example, which is accompanied by the other planets of the solar system). In fact, if anything, a single planet will be *more* stable because it will not be perturbed by other massive bodies. But in all of these cases, gravity will certainly be present. For example, the reason a planet will even stick together to *be* a planet is because gravity holds it together.

Rejected Response (Upvotes: 0)

Responding directly: Yes it is possible and is expected in cosmological time scale. It is expected for both bodies to have the same face to the other with sufficient time, but the one of smaller mass will generally slow its rotation with respect to the body of major mass first, as strictly that phenomenon is connected to the angular moment of rotation, that moment is proportional to mass and speed of rotation, but other factors influence the loss of rotation (braking), such as the existence of fluids in the celestial body (plasma in the stars and atmosphere and water or other fluid in the planets and satellites, the denser the braking of the celestial body, the energy in this case becomes tidal energy in the fluids. What speeds up the process is proximity. The body that will first present the same face is the one that has the lowest energy (moment) and the set of conditions that predispose them to lose energy faster, such as less mass, lack of atmosphere and oceans.

The earth does not yet have the same face for the Moon, but millions of years ago the day on earth lasted only about 10 hours, from here a few million years ago when the day on earth lasts for a lunar month, both the Moon and Earth will show the same face and will be synchronized. In his question the case is of the existence of a single planet in the system, this only facilitates the synchronization because there is less gravitational perturbation of other celestial bodies of great mass, but remembering that the proximity is the main factor in the speed of the process. See Wikipedia article: Gliese 581c

Search the internet for these titles for more information: Tidal coupling (Tidal Locking)

You will see that the Pluto system and its moon Charon are already synchronized each showing the same face to the other component of the system and there is a planet in the star Tau Bootes that presents this synchronization.

There is also an excellent article on the subject in the book "Asimov explains" by Isaac Asimov

Answering the second question:

Yes, the planet independent of size will have gravity, but the resulting value will be different on each side. On the side permanently presented to the star will be smaller than on the other side but will not be zero. There will be a point between the star and the planet where gravity will vanish, but it will still be off the planet. The loss of momentum of rotation partly alters the moment of translation, perhaps making it happen to the planet what already happens to our Moon that moves away from Earth at a speed of 3.78 centimeters per year, with the possibility of tearing from the earth.

Excuse my english please.

Figure 23: Example 2 drawn from StackExchange training dataset.

Query (From tex.stackexchange)

Title: How to add custom C++ keywords to be recognized by Minted?
I have defined several structures and types in a C++ project. I have to write some documentation for this code, and I use the minted package to typeset my code.
I would like to be able to add some keywords to the minted database, so minted will also recognize these and give them the appropriate color. (I would like to have them colored in the same way a type, eg. double would be colored.)
[1]: <http://www.ctan.org/pkg/minted>

Chosen Response (Upvotes: 14)

Minted uses Pygments to parse source-code, so one way to go about this is by modifying pygments/lexers/compiled.py and adding your types there (Somewhere about line 210 looks reasonable to me). This is what I had to do when Pygments was incorrectly parsing a symbol in some Erlang I'd written.
The obvious disadvantage of this is that Pygments is now non-standard. This will mean a number of things:
- This change has to be made on all PCs that want to compile this document, meaning the document will be built inconsistently or lots of people have to have modified Pygments installations. - Future updates to Pygments could break your modifications and so your document may not be correctly syntax-highlighted in the future. - You may incorrectly parse/highlight future documents if you use these keywords in a different way (So it may print a variable name in the same way it prints a type, for example).

Rejected Response (Upvotes: 4)

Another (obvious) solution would be to write your own lexer (i.e. a superset of c++)
[1]: <http://pygments.org/docs/lexerdevelopment/>

Figure 24: Example 3 drawn from StackExchange training dataset.

Query (From math.stackexchange)

Title: Does every continuous map from $\mathbb{Q}$ to $\mathbb{Q}$ extends continuously as a map from $\mathbb{R}$ to $\mathbb{R}$?

Given a continuous function $f : \mathbb{Q} \rightarrow \mathbb{Q}$, does there exist a continuous function $g : \mathbb{R} \rightarrow \mathbb{R}$ such that $g|_{\mathbb{Q}} = f$?

What I have no Idea about how to attempt this Question! Any suggestion will be very helpful.

Chosen Response (Upvotes: 16)

Not in general. Consider the function $f_0 : \mathbb{Q} \rightarrow \mathbb{Q}$ defined by

$$f_0(x) = \begin{cases} 0 & \text{if } x^2 < 2 \\ 1 & \text{if } x^2 > 2 \end{cases}.$$

It is easy to see that f_0 is continuous on $\mathbb{Q}$ (since $\sqrt{2} \notin \mathbb{Q}$), but if $g_0 : \mathbb{R} \rightarrow \mathbb{R}$ is an extension of f_0 to $\mathbb{R}$, then g_0 is necessarily discontinuous at $\sqrt{2}$.

It is worth noting that if $f : \mathbb{Q} \rightarrow \mathbb{Q}$ can be extended to a continuous function $g : \mathbb{R} \rightarrow \mathbb{R}$, then such an extension is unique. That is, if $g : \mathbb{R} \rightarrow \mathbb{R}$ and $h : \mathbb{R} \rightarrow \mathbb{R}$ are continuous functions such that $g(x) = h(x)$ for all $x \in \mathbb{Q}$, then $g = h$. This is because $\mathbb{Q}$ is a dense subset of $\mathbb{R}$.

Rejected Response (Upvotes: 6)

Yes, it's possible, iff f is sequentially continuous, in the sense that if a_n and b_n are two sequences of rational numbers that converge to the same real number, then $f(a_n)$ and $f(b_n)$ are two converging sequences that converge to the same real number.

If you want a characterisation that doesn't mention $\mathbb{R}$ at all, you probably have to use the notion of Cauchy: " a_n and b_n are Cauchy sequences such that $a_n - b_n$ converges to 0".

Figure 25: Example 4 drawn from StackExchange training dataset.