Preference Data Selection for Mitigating the Alignment Tax in Large Language Models

Minsu Kim[1,2*], **Jianxun Lian**[2†], **Xing Xie**[2], **Steven Euijong Whang**[1†]

[1]KAIST

[2]Microsoft Research Asia

Abstract

Aligning large language models to human preferences is crucial for real-world deployment but frequently incurs an alignment tax, leading to the catastrophic forgetting of pre-trained general capabilities. While previous works primarily frame this problem as an optimization or architectural challenge, the inherent characteristics of preference data that drive this degradation remain largely underexplored. In this paper, we propose BALIGN, a balanced data selection strategy that explicitly mitigates catastrophic forgetting while optimizing alignment efficacy. Through theoretical and empirical analyses of the preference optimization gradient, we identify three key data-centric features that dictate parameter drift: the reference model's log-probability margin, the token length difference between chosen and rejected responses, and the TF–IDF similarity to general capability corpora. By aggregating these orthogonal features into a unified composite risk score, BALIGN systematically filters out high-risk preference samples that disrupt intrinsic model parameters or provide minimal alignment utility. Extensive experiments on standard human preference datasets demonstrate that BALIGN strongly preserves foundational capabilities without compromising alignment gains, consistently achieving the optimal Pareto frontier with minimal computational overhead.

Introduction

Large language models (LLMs) pre-trained on massive corpora have become foundational building blocks for a wide range of natural language processing applications (Kaddour et al. 2023). A common practice in deploying these models is to take pre-trained LLMs as base models and fine-tune them to enhance their capabilities, such as mathematical reasoning (Shao et al. 2024), clinical reasoning (Chen et al. 2024), or code generation (Hui et al. 2024). Among such fine-tuning objectives, alignment with human preferences is one of the most critical, as it determines whether the resulting model produces safe, helpful, and human-aligned responses (Shen et al. 2023; Liu et al. 2023; Ji et al. 2023).

While alignment improves the quality of model responses on preference-related tasks, it is known to induce a phenomenon referred to as the alignment tax (Ouyang et al. 2022), in which the model loses a portion of its previously acquired general capabilities. This degradation constitutes a form of catastrophic forgetting (Kirkpatrick et al. 2016) as shown in Figure 1a. Such a decline in foundational skills limits the broader utility of LLMs, as their real-world application relies on these general capabilities. To mitigate this issue, effective alignment necessitates maintaining a balance between preserving general capabilities (i.e., stability) and integrating human-aligned preferences (i.e., plasticity). Despite the importance of this dual objective, most existing works on LLM fine-tuning have solely focused on adapting to new tasks for domain specialization (Xie et al. 2023; Xia et al. 2024; Liu, Karbasi, and Rekatsinas 2024).

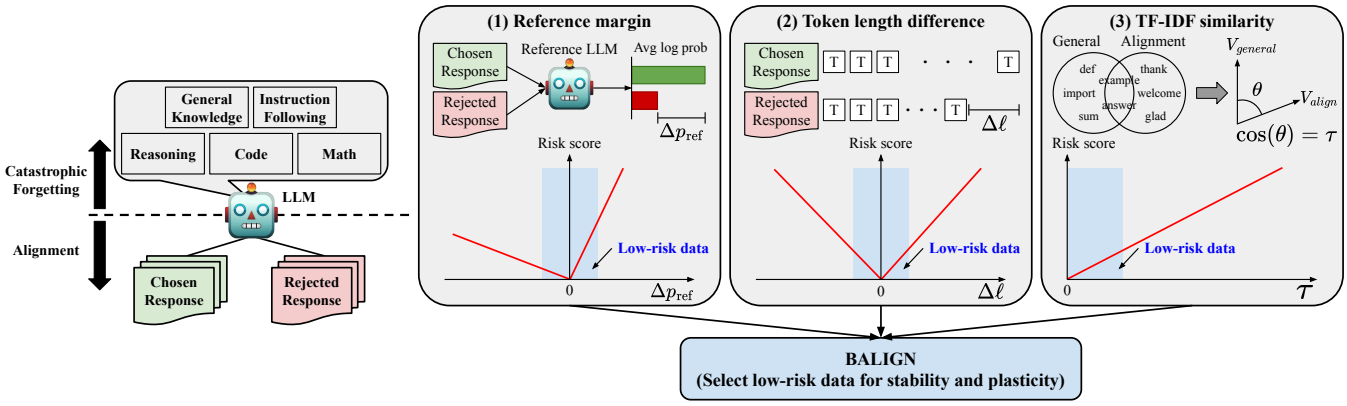
Prior work widely acknowledges that the underlying cause of catastrophic forgetting is not simply a deficit in memory capacity, but rather a disruption of intrinsic model parameters (Kirkpatrick et al. 2016; Luo et al. 2023). Based on these findings, several approaches have been proposed to mitigate catastrophic forgetting in LLM fine-tuning (Kotha, Springer, and Raghunathan 2024; Li et al. 2024; Xiong and Xie 2026). However, these methods predominantly frame catastrophic forgetting as an optimization or architectural challenge, leaving the relationship between the characteristics of new task data and catastrophic forgetting largely underexplored.

Motivated by the intuition that the fundamental difference between general and new tasks disturbs intrinsic model parameters, we focus on the data itself to analyze and mitigate catastrophic forgetting. Given that LLMs solidify the majority of their general capabilities during the SFT phase, it is crucial to analyze the interplay between the knowledge embedded in the previously learned SFT data and the learning signals provided by the newly introduced preference data from a stability-plasticity perspective. To the best of our knowledge, no existing data-centric method simultaneously optimizes both stability and plasticity in LLM alignment.

We identify three key features that characterize the impact of training samples on the stability and plasticity of LLM alignment, as shown in Figure 1b. Specifically, the length-normalized log-probability margin under the base reference model (i.e., the initial pre-trained model prior to alignment) and the token length difference between the chosen and rejected responses serve as indicators for stability, whereas the cosine similarity between the TF–IDF representations of training samples and those of a general benchmark corpus acts as a proxy for plasticity. Our theoretical and empirical analyses reveal that selecting training samples with minimal length-normalized log-probability margins and small token length differences effectively mitigates catastrophic forget-

*Work done during an internship at Microsoft Research Asia.

†Corresponding authors.



(a) (b)

Figure 1: (a) Catastrophic forgetting of general capabilities (e.g., general knowledge, instruction following, reasoning, code, and math) in LLM alignment. (b) Overview of BALIGN, which selects low-risk data for stability and plasticity using risk scores based on three features: reference margin, token length difference, and TF–IDF similarity.

ting. In addition, selecting training samples that are highly distinct from the general SFT corpus based on their TF–IDF representations yields better model alignment.

Based on our findings, we propose BALIGN, a balanced data selection strategy for effective LLM alignment that jointly addresses stability and plasticity. We first define a risk score tailored to each of the three features, namely *reference margin*, *token length difference*, and *TF–IDF similarity*, such that training on samples with high risk scores leads to degradation in either stability or plasticity. We then formulate a composite risk score that accounts for stability and plasticity by aggregating the three normalized risk scores. Based on the composite risk scores computed across all samples, we filter out high-risk samples and select a predefined proportion of low-risk samples to form the training set. Experiments on human preference datasets demonstrate that BALIGN effectively mitigates catastrophic forgetting across diverse capabilities while preserving alignment performance with a filtered data subset. We believe that our data selection strategy serves as a simple yet effective method for preprocessing training data in LLM alignment.

Summary of Contributions: (1) We identify three orthogonal data-centric features, namely reference margin, token length difference, and TF–IDF similarity, and analyze their impact on the stability-plasticity trade-off during LLM alignment; (2) We propose a balanced data selection strategy that mitigates catastrophic forgetting by filtering out high-risk preference data using a composite risk score tailored to these three features; (3) We empirically demonstrate that our approach successfully preserves general capabilities across diverse domains while maintaining alignment performance.

Related Work

Large Language Model Alignment. LLM alignment refers to the process of adjusting a pre-trained language model so that its outputs conform to human preferences regarding helpfulness, harmlessness, honesty, and safety (Shen et al. 2023; Lu et al. 2025). A foundational approach is Reinforcement Learning from Human Feedback (RLHF) (Christiano et al. 2017; Stiennon et al. 2020), which combines supervised fine-tuning with a learned reward model and Proximal Policy Optimization (PPO) (Schulman et al. 2017) to align the policy with human preferences. While recent advancements such as Group Relative Policy Optimization (GRPO) (Shao et al. 2024) mitigate the substantial memory overhead of the critic model in PPO, on-policy optimization remains engineering-heavy and computationally expensive. To remove the engineering complexity of explicit reward modeling and on-policy optimization altogether, Direct Preference Optimization (DPO) (Rafailov et al. 2023) reparameterizes the alignment objective as a single logistic loss over preference pairs, treating the policy itself as an implicit reward function. Given its simplicity and status as the standard in offline alignment (Llama Team 2024; Yang et al. 2024a), we build our alignment framework upon DPO.

Catastrophic Forgetting of Large Language Models. A well-known challenge in LLM fine-tuning is catastrophic forgetting, in which a model forgets previously acquired knowledge while learning new tasks (Kirkpatrick et al. 2016; Li et al. 2022). In LLM alignment, optimizing models for human preferences often leads to the degradation of pre-trained capabilities, a phenomenon referred to as the alignment tax (Ouyang et al. 2022). Recent studies attribute catastrophic forgetting to representational routing failures (Kotha, Springer, and Raghunathan 2024; Jain et al. 2024), convergence to sharp minima (Li et al. 2024; Watts et al. 2026), and parameter-level weight interference (Luo et al. 2023; Huang, Cheng, and Wang 2025). However, the impact of fine-tuning data itself on catastrophic forgetting remains largely underexplored in the context of LLM alignment.

Mitigating Catastrophic Forgetting of Large Language Models. Conventional continual learning techniques, including experience replay and regularization-based approaches, were first applied to the scenario of LLM fine-tuning (Scialom, Chakrabarty, and Muresan 2022; Mok et al. 2023; Lin et al. 2023). Recent studies have proposed methods specifically designed for LLMs by leveraging their internal

properties. These include self-synthesized rehearsal (Huang et al. 2024), self-distillation (Yang et al. 2024b; Shenfeld et al. 2026), and model merging (Xiao et al. 2024; Lin et al. 2024; Alexandrov et al. 2024). While these approaches have substantially advanced the trade-off between stability and plasticity, they have primarily focused on supervised fine-tuning for task-specific performance. Consequently, the joint dynamics between catastrophic forgetting and alignment remain underexplored in the data-centric literature. Our work addresses this gap by formulating a sample-level composite risk score that explicitly accounts for both objectives.

Problem Definition

Notation and Preliminaries

Let $\mathcal{X}$ denote the space of prompts and $\mathcal{Y}$ the space of responses. Let $\pi_{\text{ref}} : \mathcal{X} \to \Delta(\mathcal{Y})$ denote a pre-trained reference language model that serves as the starting point for alignment. Let $\mathcal{D} = \left\{(x_i, y_w^{(i)}, y_l^{(i)})\right\}_{i=1}^{N}$ be a human preference dataset, where $y_w^{(i)} \succ y_l^{(i)} \mid x_i$ denotes that the response $y_w^{(i)}$ is preferred over $y_l^{(i)}$ for a given prompt x_i. We refer to $y_w^{(i)}$ as the chosen response and $y_l^{(i)}$ as the rejected response. Using this preference dataset, we aim to optimize a parameterized policy $\pi_\theta : \mathcal{X} \to \Delta(\mathcal{Y})$, starting from the initial reference model $\pi_{\theta_0} = \pi_{\text{ref}}$.

We adopt DPO (Rafailov et al. 2023) as our primary alignment objective, which optimizes the policy π_θ to satisfy the observed preferences without explicitly learning a reward model. The DPO loss is given by

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}, \mathcal{D}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log \sigma\big(\beta m_\theta(x, y_w, y_l)\big)\right], \tag{1}$$

where $\sigma(\cdot)$ denotes the sigmoid function, $\beta > 0$ is a hyperparameter controlling the strength of the implicit Kullback–Leibler regularization toward π_{ref}, and

$$m_\theta(x, y_w, y_l) = \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \tag{2}$$

is the reward margin between chosen and rejected responses.

Catastrophic Forgetting and Alignment

We characterize the policy π_θ along two orthogonal axes: *general capability*, which tracks the retention of previously acquired general knowledge, and *alignment quality*, which measures adherence to human preferences.

General Capability. Let $\mathcal{E} = \{e_1, e_2, \ldots, e_K\}$ denote a suite of general-capability benchmarks that are disjoint from the alignment distribution. For each benchmark e_k, let $c_{e_k}(\pi) \in [0, 1]$ denote the normalized performance of policy π on e_k. We define the *aggregate general capability* as:

$$C(\pi) = \frac{1}{K} \sum_{k=1}^{K} c_{e_k}(\pi). \tag{3}$$

Based on this, we define the *catastrophic forgetting* of general capabilities induced by alignment as:

$$F(\pi_\theta) = C(\pi_{\text{ref}}) - C(\pi_\theta). \tag{4}$$

A positive $F(\pi_\theta)$ indicates that alignment has degraded the model's general capability relative to the reference.

Alignment Quality. Let $\mathcal{D}_{\text{test}} = \{(x_j, y_w^{(j)}, y_l^{(j)})\}_{j=1}^{M}$ denote a held-out preference test set. We measure *alignment quality* by the preference accuracy of the policy, defined as the proportion of samples where the policy assigns a higher log-probability to the chosen response $y_w^{(j)}$ than to the rejected response $y_l^{(j)}$, given by:

$$A(\pi_\theta) = \frac{1}{M} \sum_{j=1}^{M} \mathbb{1}\left[\log \pi_\theta\big(y_w^{(j)} \mid x_j\big) > \log \pi_\theta\big(y_l^{(j)} \mid x_j\big)\right]. \tag{5}$$

Building on this, we define the *alignment gain* to quantify the net improvement in alignment quality as:

$$R(\pi_\theta) = A(\pi_\theta) - A(\pi_{\text{ref}}). \tag{6}$$

Balanced Data Selection for Alignment

We frame preference data selection as a bi-level optimization problem to navigate the trade-off between alignment and catastrophic forgetting. Given a sampling ratio $\rho \in (0, 1]$, we aim to identify an optimal subset $\mathcal{S} \subseteq \mathcal{D}$ of size $\lceil \rho N \rceil$ that maximizes the joint objective:

$$\max_{\mathcal{S}\subseteq\mathcal{D},\, |\mathcal{S}|=\lceil\rho N\rceil} \Big(R(\pi_\theta^{\mathcal{S}}) - F(\pi_\theta^{\mathcal{S}})\Big), \tag{7}$$

where the inner objective defines the policy $\pi_\theta^{\mathcal{S}}$, which is optimized on the selected subset $\mathcal{S}$ via the DPO loss:

$$\pi_\theta^{\mathcal{S}} = \arg\min_{\theta\in\Theta} \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}, \mathcal{S}). \tag{8}$$

Method

We propose BALIGN, a balanced data selection strategy for LLM alignment that jointly addresses stability and plasticity. Specifically, we derive three key features that capture how preference data influences the trade-off between mitigating catastrophic forgetting and enhancing alignment quality.

Theoretical Analysis

We begin by examining the gradient of the DPO loss with respect to the policy parameters θ, which governs both the speed of alignment and the magnitude of parameter drift away from the reference π_{ref}. Differentiating Eq. (1) yields

$$\nabla_\theta \mathcal{L}_{\text{DPO}} = -\beta \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Bigg[\underbrace{\sigma\big(-\beta m_\theta(x, y_w, y_l)\big)}_{\text{per-sample weight } w_\theta} \cdot \underbrace{\big(\nabla_\theta \log \pi_\theta(y_w \mid x) - \nabla_\theta \log \pi_\theta(y_l \mid x)\big)}_{\text{per-sample direction } \Delta_\theta}\Bigg], \tag{9}$$

where each preference sample contributes a scalar weight w_θ and a direction Δ_θ to the overall parameter update. Based on this structural decomposition, the three features we propose below characterize the gradient direction Δ_θ in terms of its informational utility, accumulation scale, and support structure in the parameter space.

Reference Margin. The first feature is the reference margin Δp_{ref}, defined as the difference between the average per-token log-probabilities of the chosen and rejected responses under the reference model π_{ref}:

$$\Delta p_{\text{ref}}(x, y_w, y_l) = \frac{1}{|y_w|} \log \pi_{\text{ref}}\big(y_w \mid x\big) - \frac{1}{|y_l|} \log \pi_{\text{ref}}\big(y_l \mid x\big). \tag{10}$$

The magnitude and sign of Δp_{ref} dictate the risk of catastrophic forgetting by reflecting how the reference model inherently perceives the preference pair. Samples with a large positive Δp_{ref} indicate that the reference model already intrinsically prefers y_w over y_l. In such cases, updating the model further along their respective gradient directions fails to impart novel alignment signals and over-optimizes existing preferences, thereby amplifying representational drift with negligible learning benefit. Conversely, samples with a large negative Δp_{ref} represent cases where the reference model strongly favors y_l. Forcing the model to overturn this deeply encoded prior knowledge requires substantial parameter change, similarly driving catastrophic forgetting.

Motivated by these insights, we aim to penalize samples with a large $|\Delta p_{\text{ref}}|$ during the data selection process. To achieve this, we leverage a pinball loss formulation to define a reference margin-based risk score as follows:

Definition 1. Let $\Delta p_{\text{ref}}^{(i)} = \Delta p_{\text{ref}}(x_i, y_w^{(i)}, y_l^{(i)})$ denote the reference margin of sample i. The *reference margin-based risk score* of sample i is:

$$s_{\Delta p_{\text{ref}}}(i) = \alpha \max\big(0,\, -\Delta p_{\text{ref}}^{(i)}\big) \,+\, (1-\alpha) \max\big(0,\, \Delta p_{\text{ref}}^{(i)}\big), \tag{11}$$

where $\alpha \in [0, 1]$ is a hyperparameter designed to capture the asymmetric risk associated with the sign of Δp_{ref}.

Token Length Difference. The second feature is the token length difference between the chosen and rejected responses, defined as $\Delta\ell(y_w, y_l) = |y_w| - |y_l|$, where $|y|$ denotes the number of tokens in response y. By decomposing the log-probabilities into per-token contributions, $\log \pi_\theta(y \mid x) = \sum_{t=1}^{|y|} \log \pi_\theta(y_t \mid x, y_{<t})$, we can rewrite the reward margin in Eq. (2) as

$$m_\theta(x, y_w, y_l) = |y_w| \overline{r}_\theta(y_w \mid x) - |y_l| \overline{r}_\theta(y_l \mid x), \tag{12}$$

where $\overline{r}_\theta(y \mid x) = \frac{1}{|y|} \sum_{t=1}^{|y|} \log \frac{\pi_\theta(y_t|x, y_{<t})}{\pi_{\text{ref}}(y_t|x, y_{<t})}$ is the per-token log-ratio averaged over the response. Letting $\overline{r}_w$ and $\overline{r}_l$ denote the per-token ratios, and defining $\overline{r} = \frac{1}{2}(\overline{r}_w + \overline{r}_l)$ and $\overline{\ell} = \frac{1}{2}(|y_w| + |y_l|)$, we can decompose the margin:

$$m_\theta = \overline{r} \cdot \Delta\ell + (\overline{r}_w - \overline{r}_l) \cdot \overline{\ell}. \tag{13}$$

Eq. (13) shows that the margin decomposes into a length-driven term proportional to $\Delta\ell$ and a quality-driven term proportional to the per-token ratio gap $\overline{r}_w - \overline{r}_l$. Whenever $|\Delta\ell|$ is large, the margin signal becomes dominated by sheer response length rather than by genuine preference quality, and the resulting gradient direction Δ_θ in Eq. (9) is correspondingly amplified by the total token count. As parameter drift scales with the cumulative gradient norm, large length asymmetry inflates the alignment update without conveying additional preference information, providing a theoretical pathway from $\Delta\ell$ to catastrophic forgetting. To account for this length-induced risk during data selection, we introduce the length difference-based risk score as follows:

Definition 2. Let $\Delta\ell_i = |y_w^{(i)}| - |y_l^{(i)}|$ denote the token-level length difference between the chosen and rejected responses of sample i. The *length difference-based risk score* of sample i is defined as:

$$s_{\Delta\ell}(i) = \gamma \max\big(0,\, -\Delta\ell_i\big) \,+\, (1-\gamma) \max\big(0,\, \Delta\ell_i\big), \tag{14}$$

where $\gamma \in [0, 1]$ is a hyperparameter designed to capture the asymmetric risk associated with the sign of $\Delta\ell$.

TF–IDF Similarity. The third feature is the TF–IDF similarity, which measures how closely the lexical distribution of each preference sample aligns with that of the general capability datasets we aim to preserve. Given a corpus $\mathcal{G}$ comprising a subset of general capability datasets, we fit a TF–IDF vectorizer on the corpus $\mathcal{D} \cup \mathcal{G}$ and project both preference and general data into the same vector space. To align the format between preference and general data, we compute the TF–IDF vector of each preference sample using its prompt and chosen response. Using these vector representations, the TF–IDF similarity of sample i is defined as the average cosine similarity to the general capability corpus:

$$\tau_i \;=\; \frac{1}{|\mathcal{G}|} \sum_{(x_j, y_j) \in \mathcal{G}} \frac{\big\langle \mathbf{v}(x_i, y_w^{(i)}),\, \mathbf{v}(x_j, y_j) \big\rangle}{\big\|\mathbf{v}(x_i, y_w^{(i)})\big\| \, \big\|\mathbf{v}(x_j, y_j)\big\|}, \tag{15}$$

where $\mathbf{v}(x_i, y_w^{(i)})$ and $\mathbf{v}(x_j, y_j)$ denote the vector representations of the preference sample and the general sample, respectively. We note that since all components of a TF–IDF vector are non-negative, the resulting cosine similarity τ_i is inherently non-negative and bounded within $[0, 1]$.

When the similarity τ_i is high, the preference and general samples share an extensive vocabulary, meaning that their gradients actively update overlapping parameter subspaces. However, when τ_i is low, the preference sample introduces distinct lexical structures, such as novel instructional formats or specific alignment markers, that are largely absent from $\mathcal{G}$. Then the gradients of the preference and general samples are supported on substantially disjoint parameter regions, rendering them nearly orthogonal. As these updates do not conflict, the model can fully absorb the alignment signals without structural resistance from pre-trained priors. This orthogonality ensures that the optimization step maximizes alignment efficacy while leaving the general capability intact. Since higher similarity corresponds to a greater risk of gradient interference, we define the similarity-based risk score as the similarity value itself:

Definition 3. Let $\tau_i \in [0, 1]$ denote the TF–IDF similarity of sample i. The *similarity-based risk score* of sample i is

$$s_\tau(i) \;=\; \tau_i. \tag{16}$$

Empirical Analysis

To verify the theoretical analysis and derived scores, we conduct a bucket-based experiment in which each of the three features is used in isolation to partition the preference

Bucket	$s_{\Delta p_{\text{ref}}}$		$s_{\Delta\ell}$		s_τ	
	F ↓	R ↑	F ↓	R ↑	F ↓	R ↑
$\mathcal{B}^1$ (lowest risk)	0.08	3.33	0.38	3.20	0.95	3.80
$\mathcal{B}^2$	0.37	3.46	0.82	2.95	1.43	3.93
$\mathcal{B}^3$	0.52	4.02	1.77	2.26	0.73	3.42
$\mathcal{B}^4$	0.84	3.33	1.35	3.20	0.84	2.61
$\mathcal{B}^5$ (highest risk)	3.48	1.75	2.19	3.12	0.85	1.67

Table 1: Analysis of catastrophic forgetting F (lower is better) and alignment gain R (higher is better) for each risk score.

dataset into buckets. We use the HH-RLHF (Bai et al. 2022) dataset comprising preference triples, and adopt Llama-3.1-8B-Instruct (Llama Team 2024) as the reference model. For each feature $f \in \{\Delta p_{\text{ref}}, \Delta\ell, \tau\}$, we compute the corresponding risk score for every sample and then partition the dataset into five equal-sized quantile buckets $\{\mathcal{B}_f^1, \ldots, \mathcal{B}_f^5\}$, ordered from the lowest to the highest risk score. We set the hyperparameters $\alpha = 0.25$ and $\gamma = 0.5$ for the reference margin and length difference risk scores, respectively. These values reflect our empirical observations that the reference margin exhibits asymmetry where positive values incur greater risk, whereas the length difference demonstrates symmetry as positive and negative values offset each other.

Using each bucket, we train a separate policy and evaluate its alignment quality and the extent of catastrophic forgetting on a general benchmark suite, including MMLU (Hendrycks et al. 2021), IFEval (Zhou et al. 2023), ARC-Challenge (Clark et al. 2018), HumanEval (Chen et al. 2021), and GSM8K (Cobbe et al. 2021). We report the catastrophic forgetting (F) and alignment gain (R) across five buckets for each risk score as shown in Table 1. The results show that our proposed risk scores capture complementary dimensions of model degradation. Specifically, $s_{\Delta p_{\text{ref}}}$ and $s_{\Delta\ell}$ serve as primary indicators of catastrophic forgetting, while s_τ captures the failure of preference learning.

BALIGN Data Selection Strategy

Motivated by the complementary roles of the three risk scores, we propose a unified data selection strategy, BALIGN. To jointly mitigate catastrophic forgetting and maximize alignment gain, we place the individual risk scores on a common scale and aggregate them into a single composite metric, formally defined as follows:

Definition 4. The *composite risk score* of sample i is defined as the weighted sum of the three normalized risk scores:

$$s(i) = \widetilde{s}_{\Delta p_{\text{ref}}}(i) + \widetilde{s}_{\Delta\ell}(i) + \lambda\, \widetilde{s}_\tau(i), \tag{17}$$

where $\widetilde{s}_f(i) \in [0, 1]$ is the min–max normalized risk score for each feature $f \in \{\Delta p_{\text{ref}}, \Delta\ell, \tau\}$, and λ is a hyperparameter that controls the balance between the risk scores.

The composite score $s(i)$ thus integrates the dual aspects of stability and plasticity into a single scalar that quantifies the total expected risk of including sample i in preference learning. Following a score-based data selection template to minimize risks, our selected subset under a budget ratio $\rho \in (0, 1]$ is given by

$$\mathcal{S} = \big\{(x_i, y_w^{(i)}, y_l^{(i)}) \in \mathcal{D} : s(i) \le q_\rho(s)\big\}, \tag{18}$$

where $q_\rho(s)$ is the ρ-quantile of the composite scores $\{s(i)\}_{i=1}^N$. The final aligned policy is obtained by solving

$$\pi_\theta^{\mathcal{S}} = \arg\min_{\theta\in\Theta} \mathcal{L}_{\text{DPO}}\big(\pi_\theta; \pi_{\text{ref}}, \mathcal{S}\big). \tag{19}$$

Experiments

We conduct experiments to evaluate BALIGN and provide a detailed analysis of the results. We provide more detailed experimental settings and results in Appendix.

Metrics. We evaluate the methods along two axes: general capability and alignment, which correspond to stability and plasticity, respectively. We first measure the average absolute performance of the models on general capability ($General$) and alignment ($Align$) using Eqs. (3) and (5), respectively. We then quantify the performance shifts relative to the base reference model, defining these differences as catastrophic forgetting (F) and alignment gain (R) using Eqs. (4) and (6).

Datasets. For alignment, we use the HH-RLHF (Bai et al. 2022) dataset to train the model for helpfulness and harmlessness. We utilize separate training and test splits of the dataset for training and evaluation. For general capability, we use a total of five datasets following the benchmark evaluation settings (Jiang et al. 2023; Llama Team 2024; Yang et al. 2024a), including MMLU (Hendrycks et al. 2021), IFEval (Zhou et al. 2023), ARC-Challenge (Clark et al. 2018), HumanEval (Chen et al. 2021), and GSM8K (Cobbe et al. 2021). These datasets cover general knowledge, instruction following, reasoning, code, and math, which constitute the fundamental capabilities of LLMs. We reserve a subset of the general capability datasets as a validation set for required baselines, while the remainder serves as a test set.

Models. We adopt Llama-3.1-8B-Instruct (Llama Team 2024) and Qwen2.5-7B-Instruct (Yang et al. 2024a) as the base reference models for our experiments. We select instruction-tuned models in the 7B–8B parameter range as they offer an optimal balance between robust base capabilities and tractability within our GPU computational budget.

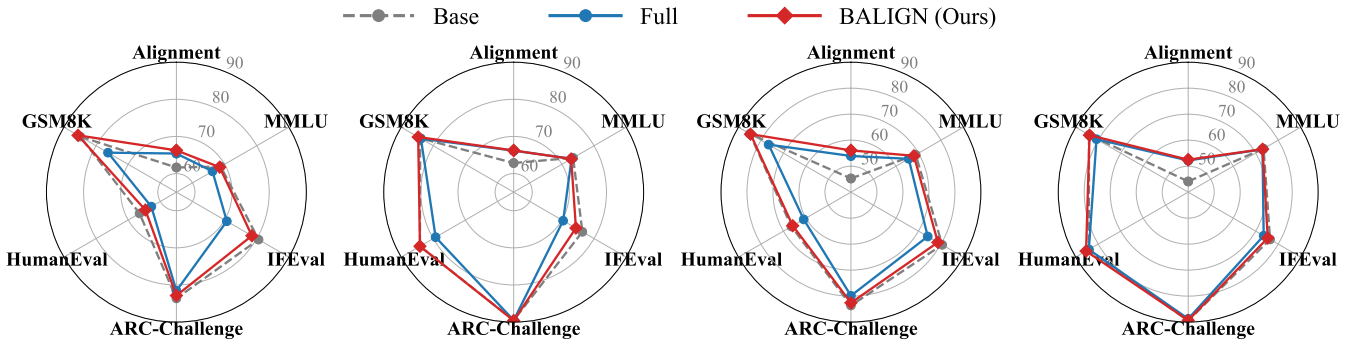
Baselines. We compare BALIGN with four categories of data selection baselines. Within this taxonomy, BALIGN is a stability- and plasticity-aware alignment method.

- **Naive methods**: *Full* and *Random*.
- **Plasticity-aware SFT methods**: *DSIR* (Xie et al. 2023), *LESS* (Xia et al. 2024), *TSDS* (Liu, Karbasi, and Rekatsinas 2024), and *NICE* (Wang et al. 2025).
- **Stability- and Plasticity-aware SFT methods**: *GrADS* (Liu et al. 2025) and *OGS* (Zhang et al. 2026b).
- **Plasticity-aware Alignment methods**: *Selective DPO* (Gao et al. 2025), *BeeS* (Deng et al. 2025), and *PD* (Zhang et al. 2026a).

Hyperparameters. We perform a grid search on the held-out validation set to find the optimal hyperparameters. For the reference margin-based risk score, we set α to 0.25 and 0.15 for the Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct models, respectively. For the length difference-based risk score, we set γ to 0.5 and 0.25, respectively. When formulating the composite risk score, we set λ to 1 and 2, respectively.

Method	Helpfulness				Harmlessness			
	Llama-3.1-8B-Instruct		Qwen2.5-7B-Instruct		Llama-3.1-8B-Instruct		Qwen2.5-7B-Instruct	
	General (F↓)	Align (R↑)	General (F↓)	Align (R↑)	General (F↓)	Align (R↑)	General (F↓)	Align (R↑)
Base	76.86 (–)	61.64 (–)	81.52 (–)	62.84 (–)	76.86 (–)	45.25 (–)	81.52 (–)	44.11 (–)
Full	71.51 (5.35↓)	65.40 (3.76↑)	79.21 (2.31↓)	66.17 (3.33↑)	71.44 (5.42↓)	53.92 (8.67↑)	80.03 (1.49↓)	52.17 (8.06↑)
Random	74.77 (2.09↓)	64.80 (3.16↑)	80.39 (1.13↓)	64.76 (1.92↑)	75.78 (1.08↓)	52.25 (7.00↑)	80.82 (0.70↓)	47.57 (3.46↑)
DSIR	74.89 (1.97↓)	63.44 (1.80↑)	80.30 (1.22↓)	64.50 (1.66↑)	75.10 (1.76↓)	50.77 (5.52↑)	80.86 (0.66↓)	47.83 (3.72↑)
LESS	72.18 (4.68↓)	64.14 (2.50↑)	79.84 (1.68↓)	63.61 (0.77↑)	74.84 (2.02↓)	49.19 (3.94↑)	79.93 (1.59↓)	48.62 (4.51↑)
TSDS	75.60 (1.26↓)	64.56 (2.92↑)	80.42 (1.10↓)	64.55 (1.71↑)	75.09 (1.77↓)	50.63 (5.38↑)	80.76 (0.76↓)	46.96 (2.85↑)
NICE	73.18 (3.68↓)	65.05 (3.41↑)	80.58 (0.94↓)	64.08 (1.24↑)	74.77 (2.09↓)	49.58 (4.33↑)	79.43 (2.09↓)	48.10 (3.99↑)
GrADS	74.49 (2.37↓)	65.01 (3.37↑)	80.22 (1.30↓)	64.08 (1.24↑)	75.68 (1.18↓)	51.77 (6.52↑)	80.79 (0.73↓)	49.41 (5.30↑)
OGS	75.79 (1.07↓)	63.73 (2.09↑)	80.69 (0.83↓)	62.58 (-0.26↑)	75.24 (1.62↓)	51.42 (6.17↑)	79.49 (2.03↓)	48.05 (3.94↑)
Selective DPO	71.68 (5.18↓)	**67.28 (5.64↑)**	79.33 (2.19↓)	66.21 (3.37↑)	71.45 (5.41↓)	55.78 (10.53↑)	78.10 (3.42↓)	**53.02 (8.91↑)**
BeeS	73.32 (3.54↓)	66.00 (4.36↑)	80.28 (1.24↓)	65.70 (2.86↑)	73.86 (3.00↓)	54.62 (9.37↑)	80.39 (1.13↓)	50.63 (6.52↑)
PD	72.67 (4.19↓)	66.25 (4.61↑)	80.07 (1.45↓)	**66.30 (3.46↑)**	73.60 (3.26↓)	54.97 (9.72↑)	80.80 (0.72↓)	49.54 (5.43↑)
BALIGN (Ours)	**76.00 (0.86↓)**	66.25 (4.61↑)	**81.14 (0.38↓)**	66.23 (3.39↑)	**75.93 (0.93↓)**	**56.07 (10.82↑)**	**81.39 (0.13↓)**	52.48 (8.37↑)

Table 2: General capability and alignment results on the HH-RLHF dataset, evaluated separately for helpfulness and harmlessness. We use Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct models as base reference models. The best and second-best results are highlighted in bold and underline, respectively, excluding the base model performance.



(a) Helpfulness – Llama. (b) Helpfulness – Qwen. (c) Harmlessness – Llama. (d) Harmlessness – Qwen.

Figure 2: Detailed general capabilities and alignment of LLMs. BALIGN successfully preserves the base model's diverse general capabilities while achieving alignment performance on par with training on the full preference dataset.

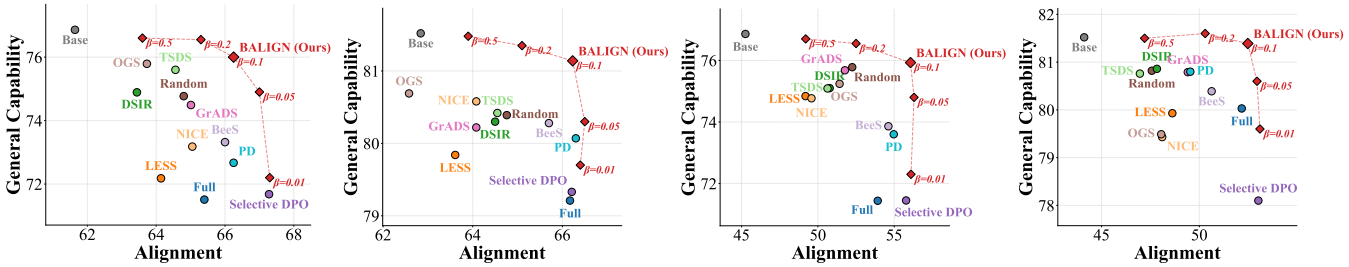
General Capability and Alignment Results

We compare BALIGN against four different categories of data selection baselines as shown in Table 2. We also present detailed capability results for each task in Figure 2 and Appendix. To ensure a fair comparison, all methods perform data selection subject to a fixed data budget, defined by a given data sampling ratio. Following standard practice in prior LLM data selection studies (Xia et al. 2024; Liu, Karbasi, and Rekatsinas 2024; Wang et al. 2025), we set the default data sampling ratio to 5% of the full dataset.

Overall, BALIGN consistently achieves the optimal balance between general capability and alignment by minimizing catastrophic forgetting while preserving alignment signals. Specifically, BALIGN achieves the best performance in general capability and the best or second-best performance in alignment, with only a marginal gap compared to the top performer. Since plasticity-aware alignment methods focus solely on alignment without considering catastrophic forgetting, outperforming them on the alignment axis is inherently difficult. In contrast, as the other data selection baselines are designed primarily for SFT data, they exhibit limitations when applied to alignment setups with preference data. These results demonstrate that BALIGN serves as a highly effective stability- and plasticity-aware alignment method.

Controllability of Stability and Plasticity

We next investigate the controllability of BALIGN by leveraging the hyperparameter β in the DPO objective, which governs the strength of the implicit Kullback-Leibler (KL) divergence penalty relative to the base reference model. By systematically varying $\beta \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$, we analyze how effectively BALIGN allows practitioners to modulate the balance between stability and plasticity as shown in Figure 3. Overall, reducing β diminishes the regularization effect, thereby generally prioritizing plasticity, whereas increasing β enforces stronger regularization to preserve stability. Across this spectrum, BALIGN consistently dominates the baselines and traces the Pareto frontier, distinctly occupying the optimal top-right region of the trade-off space, with the default β=0.1 providing a strong balance.



(a) Helpfulness – Llama. (b) Helpfulness – Qwen. (c) Harmlessness – Llama. (d) Harmlessness – Qwen.

Figure 3: Pareto frontier of BALIGN against baselines on the HH-RLHF dataset.

Method	Helpfulness				Harmlessness			
	Llama-3.1-8B-Instruct		Qwen2.5-7B-Instruct		Llama-3.1-8B-Instruct		Qwen2.5-7B-Instruct	
	General (F↓)	Align (R↑)	General (F↓)	Align (R↑)	General (F↓)	Align (R↑)	General (F↓)	Align (R↑)
Base	76.86 (–)	61.64 (–)	81.52 (–)	62.84 (–)	76.86 (–)	45.25 (–)	81.52 (–)	44.11 (–)
W/o $s_{\Delta p_{\text{ref}}}$	75.18 (1.68↓)	65.01 (3.37↑)	80.50 (1.02↓)	65.34 (2.50↑)	74.67 (2.19↓)	53.04 (7.79↑)	79.94 (1.58↓)	48.70 (4.59↑)
W/o $s_{\Delta \ell}$	75.59 (1.27↓)	65.95 (4.31↑)	79.98 (1.54↓)	65.56 (2.72↑)	74.27 (2.59↓)	53.34 (8.09↑)	80.19 (1.33↓)	51.07 (6.96↑)
W/o s_{τ}	75.78 (1.08↓)	64.03 (2.39↑)	80.44 (1.08↓)	64.89 (2.05↑)	75.30 (1.56↓)	51.11 (5.86↑)	80.80 (0.72↓)	47.96 (3.85↑)
BALIGN (Ours)	**76.00 (0.86↓)**	**66.25 (4.61↑)**	**81.14 (0.38↓)**	**66.23 (3.39↑)**	**75.93 (0.93↓)**	**56.07 (10.82↑)**	**81.39 (0.13↓)**	**52.48 (8.37↑)**

Table 3: Ablation study of BALIGN on the HH-RLHF dataset, evaluated separately for helpfulness and harmlessness.

Ablation Study

To verify the effectiveness of each risk score in BALIGN, we perform an ablation study as shown in Table 3. We evaluate the impact of individual components by removing each of the three risk scores from the composite risk score. The results demonstrate that the exclusion of any single risk score consistently compromises the equilibrium between model stability and plasticity. For example, omitting either $s_{\Delta p_{\text{ref}}}$ or $s_{\Delta \ell}$ exacerbates catastrophic forgetting, leading to a pronounced deterioration in general capabilities. Conversely, excluding s_{τ} substantially impedes the model's capacity to maximize alignment gains. These findings substantiate that the three risk scores operate in a complementary manner, exhibiting a strong synergistic effect upon integration.

Computation Time Analysis

To demonstrate the efficiency of our approach, we conduct a comparative analysis of the computational overhead incurred during the data selection phase. Specifically, we measure and compare the total computation time required by BALIGN and the baseline methods to select the optimal subset from the full preference dataset as shown in Figure 4. Gradient-based methods (e.g., LESS, NICE, OGS) and alignment methods (e.g., Selective DPO, BeeS, PD) incur prohibitive computational overhead due to expensive per-sample gradient computations or auxiliary model training. Meanwhile, GrADS achieves computational efficiency by compressing per-sample gradients into scalar norms and operating without validation targets. While statistical baselines (e.g., DSIR, TSDS) are highly efficient, they inherently ignore model-specific preference signals, yielding suboptimal alignment gains. In contrast, BALIGN effectively bridges this gap by computing three lightweight risk scores that require at most a single forward pass over the base reference model.

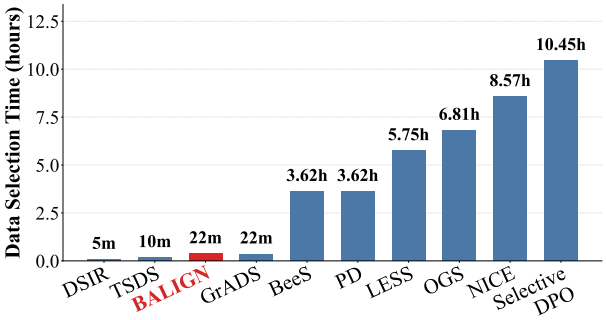


Figure 4: Computation time of data selection baselines.

Conclusion

In this work, we address the critical stability-plasticity dilemma in LLM alignment by introducing BALIGN, a highly efficient, data-centric framework designed to jointly optimize model stability and plasticity. By analyzing the preference optimization gradient, we demonstrate that catastrophic forgetting is exacerbated by preference samples exhibiting extreme reference margins and significant length asymmetries, whereas lexical overlap with general capability domains governs the model's capacity to maximize alignment gains. BALIGN seamlessly integrates these insights into a unified composite risk score, enabling the effective removal of high-risk samples prior to optimization. Our comprehensive evaluations confirm that this balanced data selection strategy not only establishes a superior trade-off between capability retention and preference alignment, but also incurs a significantly lower computational overhead than existing baselines. Therefore, BALIGN provides a scalable and practical foundation for preprocessing preference data, paving the way for developing aligned LLMs without compromising their foundational versatility.

References

Alexandrov, A.; Raychev, V.; Müller, M. N.; Zhang, C.; Vechev, M. T.; and Toutanova, K. 2024. Mitigating Catastrophic Forgetting in Language Transfer via Model Merging. In *EMNLP (Findings)*, volume EMNLP 2024 of *Findings of ACL*, 17167–17186. Association for Computational Linguistics.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; Joseph, N.; Kadavath, S.; Kernion, J.; Conerly, T.; Showk, S. E.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Hume, T.; Johnston, S.; Kravec, S.; Lovitt, L.; Nanda, N.; Olsson, C.; Amodei, D.; Brown, T. B.; Clark, J.; McCandlish, S.; Olah, C.; Mann, B.; and Kaplan, J. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *CoRR*, abs/2204.05862.

Chen, J.; Cai, Z.; Ji, K.; Wang, X.; Liu, W.; Wang, R.; Hou, J.; and Wang, B. 2024. HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs. *CoRR*, abs/2412.18925.

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; de Oliveira Pinto, H. P.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; Ray, A.; Puri, R.; Krueger, G.; Petrov, M.; Khlaaf, H.; Sastry, G.; Mishkin, P.; Chan, B.; Gray, S.; Ryder, N.; Pavlov, M.; Power, A.; Kaiser, L.; Bavarian, M.; Winter, C.; Tillet, P.; Such, F. P.; Cummings, D.; Plappert, M.; Chantzis, F.; Barnes, E.; Herbert-Voss, A.; Guss, W. H.; Nichol, A.; Paino, A.; Tezak, N.; Tang, J.; Babuschkin, I.; Balaji, S.; Jain, S.; Saunders, W.; Hesse, C.; Carr, A. N.; Leike, J.; Achiam, J.; Misra, V.; Morikawa, E.; Radford, A.; Knight, M.; Brundage, M.; Murati, M.; Mayer, K.; Welinder, P.; McGrew, B.; Amodei, D.; McCandlish, S.; Sutskever, I.; and Zaremba, W. 2021. Evaluating Large Language Models Trained on Code. *CoRR*, abs/2107.03374.

Christiano, P. F.; Leike, J.; Brown, T. B.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep Reinforcement Learning from Human Preferences. In *NIPS*, 4299–4307.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *CoRR*, abs/1803.05457.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. *CoRR*, abs/2110.14168.

Deng, X.; Zhong, H.; Ai, R.; Feng, F.; Wang, Z.; and He, X. 2025. Less is More: Improving LLM Alignment via Preference Data Selection. In *NeurIPS*.

Gao, C.; Li, H.; Liu, L.; Xie, Z.; Zhao, P.; and Xu, Z. 2025. Principled Data Selection for Alignment: The Hidden Risks of Difficult Examples. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. In *ICLR*. OpenReview.net.

Huang, J.; Cui, L.; Wang, A.; Yang, C.; Liao, X.; Song, L.; Yao, J.; and Su, J. 2024. Mitigating Catastrophic Forgetting in Large Language Models with Self-Synthesized Rehearsal. In *ACL (1)*, 1416–1428. Association for Computational Linguistics.

Huang, W.; Cheng, A.; and Wang, Y. 2025. Mitigating Catastrophic Forgetting in Large Language Models with Forgetting-aware Pruning. In *EMNLP*, 21842–21856. Association for Computational Linguistics.

Hui, B.; Yang, J.; Cui, Z.; Yang, J.; Liu, D.; Zhang, L.; Liu, T.; Zhang, J.; Yu, B.; Dang, K.; Yang, A.; Men, R.; Huang, F.; Ren, X.; Ren, X.; Zhou, J.; and Lin, J. 2024. Qwen2.5-Coder Technical Report. *CoRR*, abs/2409.12186.

Jain, S.; Kirk, R.; Lubana, E. S.; Dick, R. P.; Tanaka, H.; Rocktäschel, T.; Grefenstette, E.; and Krueger, D. S. 2024. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *ICLR*. OpenReview.net.

Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; Zeng, F.; Ng, K. Y.; Dai, J.; Pan, X.; O'Gara, A.; Lei, Y.; Xu, H.; Tse, B.; Fu, J.; McAleer, S.; Yang, Y.; Wang, Y.; Zhu, S.; Guo, Y.; and Gao, W. 2023. AI Alignment: A Comprehensive Survey. *CoRR*, abs/2310.19852.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de Las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. *CoRR*, abs/2310.06825.

Kaddour, J.; Harris, J.; Mozes, M.; Bradley, H.; Raileanu, R.; and McHardy, R. 2023. Challenges and Applications of Large Language Models. *CoRR*, abs/2307.10169.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N. C.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; Hassabis, D.; Clopath, C.; Kumaran, D.; and Hadsell, R. 2016. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796.

Kotha, S.; Springer, J. M.; and Raghunathan, A. 2024. Understanding Catastrophic Forgetting in Language Models via Implicit Inference. In *ICLR*. OpenReview.net.

Li, D.; Chen, Z.; Cho, E.; Hao, J.; Liu, X.; Xing, F.; Guo, C.; and Liu, Y. 2022. Overcoming Catastrophic Forgetting During Domain Adaptation of Seq2seq Language Generation. In *NAACL-HLT*, 5441–5454. Association for Computational Linguistics.

Li, H.; Ding, L.; Fang, M.; and Tao, D. 2024. Revisiting Catastrophic Forgetting in Large Language Model Tuning. In *EMNLP (Findings)*, volume EMNLP 2024 of *Findings of ACL*, 4297–4308. Association for Computational Linguistics.

Lin, Y.; Lin, H.; Xiong, W.; Diao, S.; Liu, J.; Zhang, J.; Pan, R.; Wang, H.; Hu, W.; Zhang, H.; Dong, H.; Pi, R.; Zhao, H.; Jiang, N.; Ji, H.; Yao, Y.; and Zhang, T. 2024. Mitigating the Alignment Tax of RLHF. In *EMNLP*, 580–606. Association for Computational Linguistics.

Lin, Y.; Tan, L.; Lin, H.; Zheng, Z.; Pi, R.; Zhang, J.; Diao, S.; Wang, H.; Zhao, H.; Yao, Y.; et al. 2023. Speciality vs generality: An empirical study on catastrophic forgetting in fine-tuning foundation models. *arXiv preprint arXiv:2309.06256*, 11: 14.

Liu, Y.; Wang, S.; Liu, Z.; Song, Z.; Wang, J.; Liu, J.; Liu, Q.; and Wang, Y. 2025. Learn More, Forget Less: A Gradient-Aware Data Selection Approach for LLM. *CoRR*, abs/2511.08620.

Liu, Y.; Yao, Y.; Ton, J.; Zhang, X.; Guo, R.; Cheng, H.; Klochkov, Y.; Taufiq, M. F.; and Li, H. 2023. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. *CoRR*, abs/2308.05374.

Liu, Z.; Karbasi, A.; and Rekatsinas, T. 2024. TSDS: Data Selection for Task-Specific Model Finetuning. In *NeurIPS*.

Llama Team. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.

Lu, H.; Fang, L.; Zhang, R.; Li, X.; Cai, J.; Cheng, H.; Tang, L.; Liu, Z.; Sun, Z.; Wang, T.; Zhang, Y.; Zidan, A. H.; Xu, J.; Yu, J.; Yu, M.; Jiang, H.; Gong, X.; Luo, W.; Sun, B.; Chen, Y.; Ma, T.; Wu, S.; Zhou, Y.; Chen, J.; Xiang, H.; Zhang, J.; Jahin, A.; Ruan, W.; Deng, K.; Pan, Y.; Wang, P.; Li, J.; Liu, Z.; Zhang, L.; Zhao, L.; Liu, W.; Zhu, D.; Xing, X.; Dou, F.; Zhang, W.; Huang, C.; Liu, R.; Zhang, M.; Liu, Y.; Sun, X.; Lu, Q.; Xiang, Z.; Zhong, W.; Liu, T.; and Ma, P. 2025. Alignment and Safety in Large Language Models: Safety Mechanisms, Training Paradigms, and Emerging Challenges. *CoRR*, abs/2507.19672.

Luo, Y.; Yang, Z.; Meng, F.; Li, Y.; Zhou, J.; and Zhang, Y. 2023. An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning. *CoRR*, abs/2308.08747.

Mok, J.; Do, J.; Lee, S.; Taghavi, T.; Yu, S.; and Yoon, S. 2023. Large-scale Lifelong Learning of In-context Instructions and How to Tackle It. In *ACL (1)*, 12573–12589. Association for Computational Linguistics.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *NeurIPS*.

Rajbhandari, S.; Rasley, J.; Ruwase, O.; and He, Y. 2020. ZeRO: memory optimizations toward training trillion parameter models. In *SC*, 20. IEEE/ACM.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *CoRR*, abs/1707.06347.

Scialom, T.; Chakrabarty, T.; and Muresan, S. 2022. Fine-tuned Language Models are Continual Learners. In *EMNLP*, 6107–6122. Association for Computational Linguistics.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *CoRR*, abs/2402.03300.

Shen, T.; Jin, R.; Huang, Y.; Liu, C.; Dong, W.; Guo, Z.; Wu, X.; Liu, Y.; and Xiong, D. 2023. Large Language Model Alignment: A Survey. *CoRR*, abs/2309.15025.

Shenfeld, I.; Damani, M.; Hübotter, J.; and Agrawal, P. 2026. Self-Distillation Enables Continual Learning. *CoRR*, abs/2601.19897.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D. M.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize from human feedback. *CoRR*, abs/2009.01325.

Wang, J.; Lin, X.; Qiao, R.; Koh, P. W.; Foo, C.; and Low, B. K. H. 2025. NICE Data Selection for Instruction Tuning in LLMs with Non-differentiable Evaluation Metric. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net.

Watts, I.; Li, C.; Goyal, S.; Springer, J. M.; and Raghunathan, A. 2026. Sharpness-Aware Pretraining Mitigates Catastrophic Forgetting. *CoRR*, abs/2605.02105.

Xia, M.; Malladi, S.; Gururangan, S.; Arora, S.; and Chen, D. 2024. LESS: Selecting Influential Data for Targeted Instruction Tuning. In *ICML*, volume 235 of *Proceedings of Machine Learning Research*, 54104–54132. PMLR / OpenReview.net.

Xiao, S.; Liu, Z.; Zhang, P.; and Xing, X. 2024. LM-Cocktail: Resilient Tuning of Language Models via Model Merging. In *ACL (Findings)*, volume ACL 2024 of *Findings of ACL*, 2474–2488. Association for Computational Linguistics.

Xie, S. M.; Santurkar, S.; Ma, T.; and Liang, P. 2023. Data Selection for Language Models via Importance Resampling. In *NeurIPS*.

Xiong, Y.; and Xie, X. 2026. OPLoRA: Orthogonal Projection LoRA Prevents Catastrophic Forgetting During Parameter-Efficient Fine-Tuning. In *AAAI*, 34088–34096. AAAI Press.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2024a. Qwen2.5 Technical Report. *CoRR*, abs/2412.15115.

Yang, Z.; Pang, T.; Feng, H.; Wang, H.; Chen, W.; Zhu, M.; and Liu, Q. 2024b. Self-Distillation Bridges Distribution Gap in Language Model Fine-Tuning. In *ACL (1)*, 1028–1043. Association for Computational Linguistics.

Zhang, J.; Liu, Y.; Zhang, C.-X.; Liu, Y.; Jin, Y.-X.; Guo, L.-Z.; and Li, Y.-F. 2026a. Data Selection for LLM Alignment Using Fine-Grained Preferences. arXiv:2508.07638.

Zhang, X.; Tian, Y.; Wang, H.; and Song, Y. 2026b. Training Data Selection with Gradient Orthogonality for Efficient Domain Adaptation. *CoRR*, abs/2602.06359.

Zhou, J.; Lu, T.; Mishra, S.; Brahma, S.; Basu, S.; Luan, Y.; Zhou, D.; and Hou, L. 2023. Instruction-Following Evaluation for Large Language Models. *CoRR*, abs/2311.07911.

Appendix

More Details on Empirical Analysis

We provide more details on empirical analysis in Table 1. The results demonstrate that our proposed risk scores capture complementary dimensions of model degradation. Specifically, $s_{\Delta p_{\text{ref}}}$ and $s_{\Delta \ell}$ serve as primary indicators of catastrophic forgetting. As the risk increases from $\mathcal{B}^1$ to $\mathcal{B}^5$, the forgetting metric exhibits a pronounced upward trajectory for Δp_{ref} and $\Delta \ell$, reaching 3.48% and 2.19%, respectively. In contrast, s_τ primarily captures the failure of preference learning. For this feature, the highest-risk bucket yields a significantly diminished alignment gain while maintaining a relatively stable forgetting rate. These findings empirically substantiate our theoretical formulation that $s_{\Delta p_{\text{ref}}}$ and $s_{\Delta \ell}$ effectively identify samples prone to destroying pre-trained knowledge, while s_τ isolates samples detrimental to the acquisition of human preferences. This distinct yet complementary behavior underscores the necessity of integrating all three features for a robust alignment framework.

More Details on Experimental Settings

We provide more details on experimental settings. All training and evaluation are run on Intel Xeon Silver 4210R CPUs and NVIDIA RTX A6000 GPUs.

Metrics. Since both general capability and alignment are crucial, we evaluate the overall effectiveness of each method based on its optimal balance between the two axes. We note that maximizing alignment gain (R) while minimizing catastrophic forgetting (F) represents desirable model performance. During evaluation, we employ greedy decoding for all generative benchmarks and log-likelihood scoring for multiple-choice tasks, thereby eliminating stochasticity from the reported metrics.

Datasets. For alignment, we use the HH-RLHF (Bai et al. 2022) dataset to train the model for helpfulness and harmlessness. We utilize separate training and test splits of the dataset for training and evaluation. Specifically, the helpfulness and harmlessness subsets contain 43,835 and 42,537 training samples with 2,354 and 2,312 test samples, respectively.

Models. We adopt Llama-3.1-8B-Instruct (Llama Team 2024) and Qwen2.5-7B-Instruct (Yang et al. 2024a) as the base reference models for our experiments. We select instruction-tuned models in the 7B–8B parameter range as they offer an optimal balance between robust base capabilities and tractability within our GPU computational budget. Since our methodology requires parameter updates during training, we cannot use closed-weight proprietary models.

Baselines. We compare BALIGN with four categories of data selection baselines. These categories are chosen to provide a comprehensive comparison, covering basic selection methods, objective-driven methods for supervised fine-tuning (SFT), and recent methods specifically tailored for preference alignment. To adapt the SFT-specific baselines to preference datasets, we implement their data selection process by treating the chosen response in each preference pair as the ground-truth target. Within this taxonomy, BALIGN is a stability- and plasticity-aware alignment method.

- **Naive methods**: *Full* uses the entire preference dataset without selection; and *Random* draws a uniformly random subset of the preference dataset.
- **Plasticity-aware SFT methods**: *DSIR* (Xie et al. 2023) performs importance resampling with hashed n-gram features to match the target-task distribution; *LESS* (Xia et al. 2024) selects samples whose low-rank gradient projections align with the target-task gradient direction; *TSDS* (Liu, Karbasi, and Rekatsinas 2024) selects samples whose dense embeddings collectively match the target-task distribution while penalizing redundancy; and *NICE* (Wang et al. 2025) extends LESS by casting selection as a policy-gradient optimization problem.
- **Stability- and Plasticity-aware SFT methods**: *GrADS* (Liu et al. 2025) applies kernel density estimation on embedding- and logit-gradient magnitudes to jointly favor in-distribution samples for both target and anchor; and *OGS* (Zhang et al. 2026b) retains samples whose projected gradients are orthogonal to a capability-anchor gradient.
- **Plasticity-aware Alignment methods**: *Selective DPO* (Gao et al. 2025) trains auxiliary DPO references on data partitions and retains samples with the lowest validation loss; *BeeS* (Deng et al. 2025) applies Bayesian aggregation to implicit and external reward margins for consensus-based selection; and *PD* (Zhang et al. 2026a) trains proxy reward models on multiple fine-grained sub-preferences and selects samples with the strongest cross-aspect consensus.

Hyperparameters. For DPO training, we set the effective batch size to 128, use a learning rate of 5×10^{-6} with a cosine schedule (10% warmup) and a sigmoid preference loss with β=0.1, and train for 3 epochs in `bfloat16` precision under DeepSpeed ZeRO-3 (Rajbhandari et al. 2020) with full-parameter fine-tuning.

More Details on General Capability and Alignment Results

We provide more details on general capability and alignment results as shown in Tables 4–7.

Method	MMLU	IFEval	ARC-Challenge	HumanEval	GSM8K	General	Forgetting	Alignment	Gain
Base (Llama-3.1-8B-Instruct)	68.78	80.55	83.62	66.46	84.91	76.86	–	61.64	–
Full	66.22	70.68	81.57	62.80	76.27	71.51	5.35	65.40	3.76
Random	68.22	80.20	83.11	59.15	83.17	74.77	2.09	64.80	3.16
DSIR	68.24	78.60	82.94	62.80	81.88	74.89	1.97	63.44	1.80
LESS	68.05	72.99	82.51	60.98	76.35	72.18	4.68	64.14	2.50
TSDS	68.78	79.67	82.34	64.02	83.17	75.60	1.26	64.56	2.92
NICE	67.99	78.33	81.83	57.93	79.83	73.18	3.68	65.05	3.41
GrADS	68.17	77.47	82.51	64.02	80.29	74.49	2.37	65.01	3.37
OGS	68.57	79.55	83.28	63.41	84.15	<u>75.79</u>	<u>1.07</u>	63.73	2.09
Selective DPO	67.98	68.10	83.28	59.76	79.30	71.68	5.18	**67.28**	**5.64**
BeeS	68.32	73.31	83.62	60.37	80.97	73.32	3.54	66.00	4.36
PD	68.33	74.35	83.53	57.32	79.83	72.67	4.19	<u>66.25</u>	<u>4.61</u>
BALIGN (Ours)	68.46	78.48	82.85	64.63	85.60	**76.00**	**0.86**	<u>66.25</u>	<u>4.61</u>

Table 4: Detailed general capability and alignment results for helpfulness on the HH-RLHF dataset, using Llama-3.1-8B-Instruct as the base reference model. The best and second-best results are highlighted in bold and underline, respectively, excluding the base model performance.

Method	MMLU	IFEval	ARC-Challenge	HumanEval	GSM8K	General	Forgetting	Alignment	Gain
Base (Qwen2.5-7B-Instruct)	73.49	76.37	89.51	84.15	84.08	81.52	–	62.84	–
Full	73.02	70.38	89.51	79.27	83.85	79.21	2.31	66.17	3.33
Random	73.37	74.33	89.16	81.10	84.00	80.39	1.13	64.76	1.92
DSIR	72.85	74.00	89.51	78.66	86.50	80.30	1.22	64.50	1.66
LESS	73.47	72.37	89.33	78.66	85.37	79.84	1.68	63.61	0.77
TSDS	73.34	72.55	89.25	83.54	83.40	80.42	1.10	64.55	1.71
NICE	73.86	72.85	89.42	79.27	87.49	80.58	0.94	64.08	1.24
GrADS	73.34	74.77	89.51	79.88	83.62	80.22	1.30	64.08	1.24
OGS	72.97	75.84	89.68	82.93	82.03	<u>80.69</u>	<u>0.83</u>	62.58	-0.26
Selective DPO	73.47	72.35	89.59	77.44	83.78	79.33	2.19	66.21	3.37
BeeS	73.24	72.76	89.85	80.49	85.06	80.28	1.24	65.70	2.86
PD	73.37	72.95	89.76	78.66	85.60	80.07	1.45	**66.30**	**3.46**
BALIGN (Ours)	72.89	74.36	89.59	84.15	84.69	**81.14**	**0.38**	<u>66.23</u>	<u>3.39</u>

Table 5: Detailed general capability and alignment results for helpfulness on the HH-RLHF dataset, using Qwen2.5-7B-Instruct as the base reference model. The best and second-best results are highlighted in bold and underline, respectively, excluding the base model performance.

Method	MMLU	IFEval	ARC-Challenge	HumanEval	GSM8K	General	Forgetting	Alignment	Gain
Base (Llama-3.1-8B-Instruct)	68.78	80.55	83.62	66.46	84.91	76.86	–	45.25	–
Full	65.66	74.12	79.86	60.98	76.57	71.44	5.42	53.92	8.67
Random	67.72	79.34	83.19	65.85	82.79	<u>75.78</u>	<u>1.08</u>	52.25	7.00
DSIR	67.52	77.92	83.02	65.24	81.80	75.10	1.76	50.77	5.52
LESS	67.45	74.47	83.62	69.51	79.15	74.84	2.02	49.19	3.94
TSDS	68.03	78.51	82.34	63.41	83.17	75.09	1.77	50.63	5.38
NICE	67.01	76.12	83.19	67.68	79.83	74.77	2.09	49.58	4.33
GrADS	67.69	78.61	84.13	64.02	83.93	75.68	1.18	51.77	6.52
OGS	67.60	73.64	83.45	67.68	83.85	75.24	1.62	51.42	6.17
Selective DPO	64.86	68.59	82.08	64.63	77.10	71.45	5.41	<u>55.78</u>	<u>10.53</u>
BeeS	67.54	76.49	83.53	62.80	78.92	73.86	3.00	54.62	9.37
PD	66.86	75.30	83.28	63.41	79.15	73.60	3.26	54.97	9.72
BALIGN (Ours)	67.97	78.63	82.59	65.85	84.61	**75.93**	**0.93**	**56.07**	**10.82**

Table 6: Detailed general capability and alignment results for harmlessness on the HH-RLHF dataset, using Llama-3.1-8B-Instruct as the base reference model. The best and second-best results are highlighted in bold and underline, respectively, excluding the base model performance.

Method	MMLU	IFEval	ARC-Challenge	HumanEval	GSM8K	General	Forgetting	Alignment	Gain
Base (Qwen2.5-7B-Instruct)	73.49	76.37	89.51	84.15	84.08	81.52	–	44.11	–
Full	73.00	73.53	88.74	84.15	80.74	80.03	1.49	52.17	8.06
Random	73.16	76.70	88.99	82.93	82.34	80.82	0.70	47.57	3.46
DSIR	73.23	76.55	90.02	81.10	83.40	<u>80.86</u>	<u>0.66</u>	47.83	3.72
LESS	73.47	75.51	89.42	81.10	80.14	79.93	1.59	48.62	4.51
TSDS	73.48	76.60	88.99	82.32	82.41	80.76	0.76	46.96	2.85
NICE	73.39	74.79	89.33	79.88	79.76	79.43	2.09	48.10	3.99
GrADS	72.59	77.30	88.99	85.37	79.68	80.79	0.73	49.41	5.30
OGS	72.27	74.57	89.42	83.54	77.63	79.49	2.03	48.05	3.94
Selective DPO	72.33	71.53	88.74	79.27	78.62	78.10	3.42	**53.02**	**8.91**
BeeS	72.82	75.09	89.25	81.71	83.09	80.39	1.13	50.63	6.52
PD	73.02	74.91	89.59	82.32	84.15	80.80	0.72	49.54	5.43
BALIGN (Ours)	73.05	75.34	89.33	85.37	83.85	**81.39**	**0.13**	<u>52.48</u>	<u>8.37</u>

Table 7: Detailed general capability and alignment results for harmlessness on the HH-RLHF dataset, using Qwen2.5-7B-Instruct as the base reference model. The best and second-best results are highlighted in bold and underline, respectively, excluding the base model performance.