LifeAlign: Lifelong Alignment for Large Language Models with Memory-Augmented Focalized Preference Optimization

Junsong Li[1], Jie Zhou[1]*, Bihao Zhan[1], Yutao Yang[1], Qianjun Pan[1], Shilian Chen[1], Tianyu Huai[1], Xin Li[2], Qin Chen[1], Liang He[1]

[1] School of Computer Science and Technology, East China Normal University, Shanghai
[2] Shanghai AI Laboratory
junsong-li@stu.ecnu.edu.cn, jzhou@cs.ecnu.edu.cn

Abstract

Alignment plays a crucial role in Large Language Models (LLMs) in aligning with human preferences on a specific task/domain. Traditional alignment methods suffer from catastrophic forgetting, where models lose previously learned values when adapting to new preferences or domains. We introduce LifeAlign, a novel framework for lifelong alignment that enables LLMs to maintain consistent human preference alignment across sequential learning tasks without forgetting previously learned values. Our approach consists of two key innovations. First, we propose a focalized preference optimization strategy that aligns LLMs with new preferences while preventing the erosion of alignment acquired from previous tasks. Second, we develop a short-to-long memory consolidation mechanism that merges denoised short-term preference representations into stable long-term memory using intrinsic dimensionality reduction, enabling efficient storage and retrieval of alignment patterns across diverse domains. We evaluate LifeAlign across multiple sequential alignment tasks spanning different domains and preference types. Experimental results demonstrate that our method achieves superior performance in maintaining both preference alignment quality and knowledge retention compared to existing lifelong learning approaches.

Extended version — https://arxiv.org/abs/2509.17183

1 Introduction

Aligning Large Language Models (LLMs) with human preferences has become a central challenge in modern artificial intelligence. As LLMs are increasingly deployed across diverse applications—from conversational assistants to domain-specific experts, ensuring their behavior aligns with human values is essential. Traditional approaches such as Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al. 2022), Direct Preference Optimization (DPO) (Rafailov et al. 2023), and Constitutional AI (Bai et al. 2022b) have achieved notable success in aligning models with predefined preference sets under controlled conditions. However, these methods generally assume static preferences and are tailored for single-task optimization.

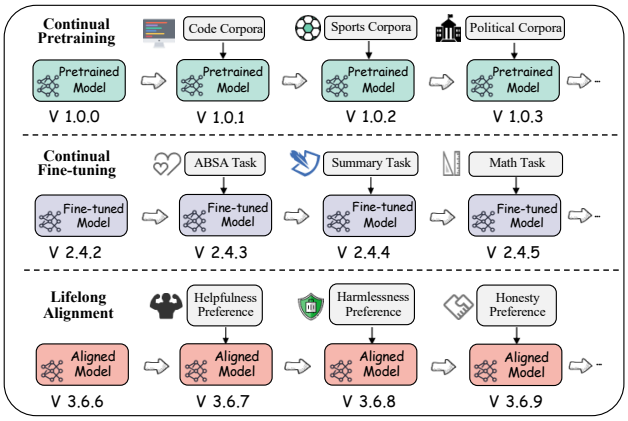

Figure 1: From Continual Pretraining, Continual Fine-Tuning to Lifelong Alignment.

In real-world deployment, LLMs face a distinct challenge: the need for lifelong alignment. As these systems operate over extended periods, they must continually adapt to evolving human preferences, new domains, and shifting societal values while preserving previously learned alignment properties. For example, a conversational AI may begin by learning to be helpful and harmless in general contexts, then adapt to specialized domains such as medicine, law, or customer service, each introducing unique preference structures. The model must sequentially acquire new alignment constraints while retaining its foundational principles. This calls for a lifelong alignment paradigm: a framework that enables LLMs to evolve with changing tasks and user expectations without compromising trustworthiness, safety, or prior alignment. Such a paradigm is vital for ensuring consistent performance, user satisfaction, and ethical reliability in real-world environments where both context and expectations are continually shifting.

Advances in lifelong learning (also known as continual learning) have sought to address the challenge of sequential task acquisition in neural networks (Wu et al. 2024; Shi et al. 2024; Yang et al. 2025a). To mitigate catastrophic forgetting, various strategies have been proposed, including regularization-based (Chaudhry et al. 2018; Wang et al. 2023a), replay-based (Rolnick et al. 2019; Lopez-Paz and Ranzato 2017), and parameter isolation-based methods

*Corresponding author.

(Wang et al. 2022), which have shown promise in computer vision and natural language processing. However, these approaches are predominantly designed for supervised learning settings and do not readily extend to the unique challenges of preference-based alignment. In such scenarios, learning signals are inherently comparative, and the optimization landscape is often non-stationary due to evolving user preferences and societal norms. In the context of LLMs, continual learning research has largely focused on two stages (as shown in Figure 1): (i) continual pre-training (CPT) and (ii) continual fine-tuning (CFT) (Wu et al. 2024; Shi et al. 2024). CPT aims to keep models up-to-date by periodically retraining them on fresh or domain-specific corpora (Ke et al. 2023; Que et al. 2024; Xie, Aggarwal, and Ahmad 2024), while CFT enhances their adaptability to new downstream tasks (Wang et al. 2023b; Jin et al. 2023). Despite this progress, relatively little attention has been paid to lifelong alignment, a critical capability that enables models to continuously refine their value judgments in response to shifting social norms and diverse, dynamic user preferences.

We introduce LifeAlign, a novel framework designed to overcome a critical limitation in LLMs: the difficulty of balancing the acquisition of new preferences with the retention of past alignment behaviors. LifeAlign is built upon two key innovations. The first, Focalized Preference Optimization (FPO), is a targeted optimization strategy that fine-tunes the model on new preferences while protecting the previously learned behaviors. The second, Short-to-Long Memory Consolidation (SLMC), is a memory-augmented mechanism that captures and compresses short-term preference representations into a stable, low-dimensional long-term memory. This module dynamically distills core alignment knowledge, suppresses conflicting signals, and seamlessly integrates the refined updates. We conducte a comprehensive empirical evaluation on a custom-built, six-task alignment dataset, demonstrating that LifeAlign achieves superior performance compared to existing lifelong learning approaches. Through LifeAlign, we take a significant step toward building LLMs that can evolve in alignment with human values over time, a key requirement for the deployment of reliable and trustworthy AI systems.

Our contributions are summarized as follows:

- We propose LifeAlign, a novel framework for lifelong alignment of LLMs that effectively addresses catastrophic forgetting in sequential alignment tasks. We formalize this critical problem and propose a solution that combines focalized optimization with memory consolidation.
- Specifically, our Focalized Preference Optimization enables targeted alignment without sacrificing performance on previous tasks and Short-to-Long Memory Consolidation consolidates stable alignment knowledge into long-term memory via intrinsic dimensionality reduction, while maintaining crucial short-term representations.
- Comprehensive experiments across diverse domains and preference types show LifeAlign's superior performance in both alignment quality and knowledge retention compared to existing lifelong learning and alignment methods.

2 Related Work

Lifelong Learning for LLMs

Lifelong learning for LLMs enables models to absorb continuous data without catastrophic forgetting. Most prior research focuses on two main stages: CPT and CFT. In CPT, models are periodically re-pretrained on updated corpora to maintain relevance, with studies such as Jang et al. (Jang et al. 2022) demonstrating incremental updates for topical relevance, and others showing improvements in domain adaptation (Ke et al. 2023; Yadav et al. 2023). In CFT, models fine-tune on new instruction-response pairs to improve task performance (Yang et al. 2025b; Huai et al. 2025), but both approaches treat alignment as a one-time process. Only CPPO (Zhang et al. 2024) and COPR (Zhang et al. 2025) explore sequential policy updates in an alignment-style setting. CPPO splits a dataset, applies supervised fine-tuning on one half and PPO on the other, and evaluates the first split. COPR extends this by using three datasets. However, both approaches are limited by a small number of tasks, which fail to capture real-world value shifts and test models against evolving preferences. Their limited number of tasks hinders the ability to test models against evolving preferences, creating a gap in lifelong alignment research.

LLMs Alignment

Alignment research for LLMs has predominantly focused on single-stage methods that align model outputs to human preferences or explicit rules. Early work, such as InstructGPT (Ouyang et al. 2022), employs reinforcement learning from human feedback (RLHF) with Proximal Policy Optimization (PPO) (Schulman et al. 2017) to steer model behavior. More recently, Direct Preference Optimization (DPO) (Rafailov et al. 2023) provides a theoretically grounded alternative that directly optimizes the preference likelihood without an explicit RL loop. Constitutional AI (Bai et al. 2022b) further augments alignment by using automated constitution checks to guide preference labeling. Furthermore, Reinforcement Learning from AI Feedback (RLAIF) reduces the reliance on costly human labels by training the reward model on preferences generated by an off-the-shelf LLM, achieving performance on a par with RLHF across summarization and dialogue tasks (Lee et al. 2024).

However, existing methods align models only once, leaving them prone to drift as preferences change. Recently, dynamic or on-the-fly adaptation of LLM behavior has emerged, focusing on principle-driven inference-time alignment to adjust model outputs based on situational values or rules. Studies (Xu et al. 2023; Zhu et al. 2025; Lu et al. 2024) incorporate normative principles into decoding for context-appropriate responses. While these methods offer flexibility, they cannot maintain alignment knowledge across tasks and mainly focus on inference-time adaptation. Thus, they complement but do not replace lifelong alignment, which preserves and adapts value alignment as preferences evolve.

Hence, we address this gap by introducing a lifelong alignment dataset and a consolidation mechanism to continuously update and refine alignment knowledge over time.

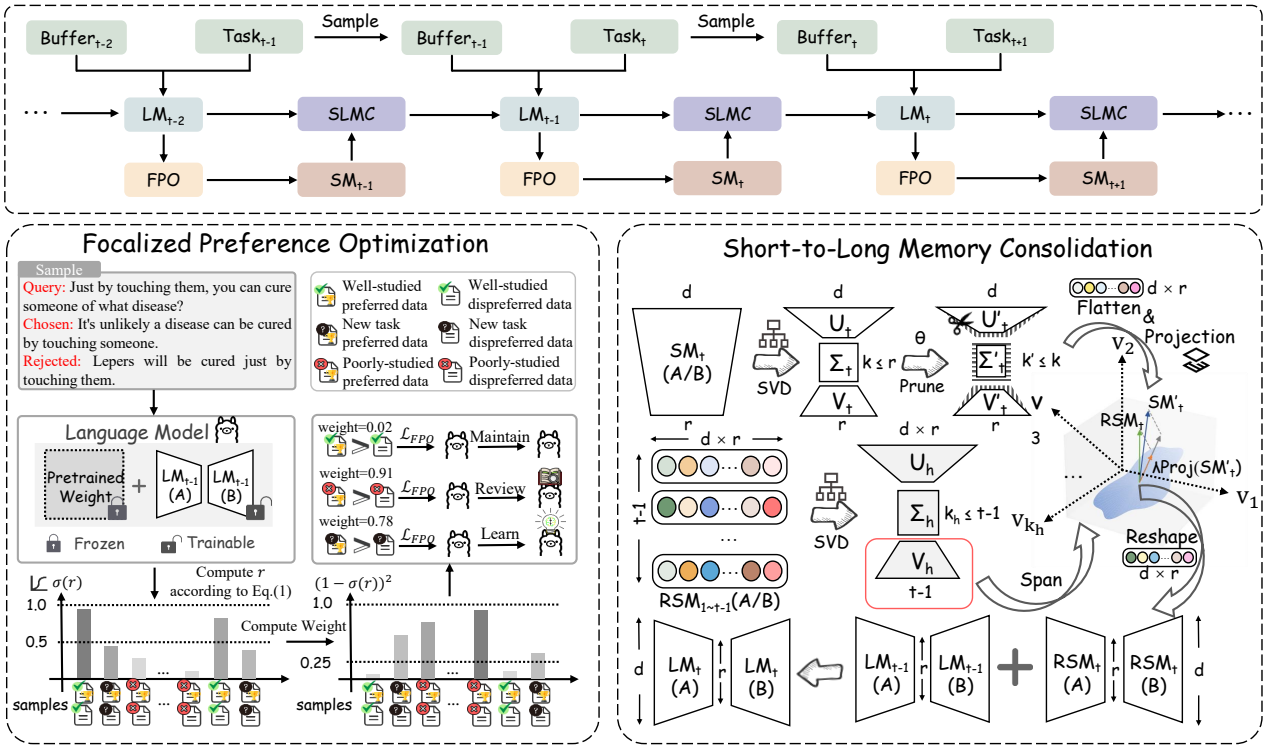

Figure 2: The overall framework of LifeAlign. LifeAlign addresses catastrophic forgetting in LLMs by enabling lifelong alignment with evolving human preferences. It integrates two core components: Focalized Preference Optimization (FPO) and Short-to-Long Memory Consolidation (SLMC). FPO (left) selectively fine-tunes the LLM on new preference data while safeguarding previously learned behaviors. SLMC (right) captures, denoises, and consolidates short-term preference representations into stable, low-dimensional long-term memory, ensuring robust retention of past alignment knowledge.

3 Our Method

LifeAlign is a novel framework designed to facilitate lifelong alignment for LLMs, enabling them to adapt to evolving human preferences across sequential tasks while mitigating catastrophic forgetting. Our method integrates two core components: Focalized Preference Optimization (FPO) and Short-to-Long Memory Consolidation (SLMC). FPO selectively fine-tunes the LLMs on new preference data, ensuring that the model learns new alignment patterns without undermining previously acquired values. Concurrently, SLMC provides a dynamic memory system that captures, denoises, and consolidates short-term preference representations into a stable, low-dimensional long-term memory.

Formally, lifelong alignment for an LLM involves sequentially learning a series of alignment preferences $\mathcal{P} = \{P_1, P_2, \ldots, P_N\}$ over a sequence of tasks $\mathcal{T} = \{T_1, T_2, \ldots, T_N\}$. For each task $T_k \in \mathcal{T}$ at time step k, the LLM is updated with a corresponding preference dataset $D_k = \{(x_i, y_i^p, y_i^d)\}_{i=1}^{M_k} \sim P_k$, where x_i is a input, y_i^p is a preferred response, and y_i^d is a dispreferred response, indicating that y_i^p is preferred over y_i^d according to preference P_k. The objective is to train a policy π_{θ_k} (parameterized by θ_k) such that, after processing all tasks up to T_k, the LLM not only achieves high alignment performance on the current task T_k (i.e., π_{θ_k} captures P_k), but also maintains high alignment performance on all previously tasks T_j for $j < k$ (i.e., π_{θ_k} retains alignment with P_j). Formally, we aim to maximize $\sum_{j=1}^{k} \mathbb{E}_{(x,y^p,y^d)\sim P_j}[\log \sigma(r_{\theta_k}(x, y^p) - r_{\theta_k}(x, y^d))]$ for $k \in \{1, \ldots, N\}$, where r_{θ_k} is the reward model derived from π_{θ_k}.

Focalized Preference Optimization

A core challenge in lifelong alignment is that standard, static loss functions are ill-suited for a dynamic learning process. They apply equal learning pressure to all examples indiscriminately, regardless of whether the examples are new, old, easy, or hard, causing the model to overwrite knowledge from past tasks that was already well-established. To endow an LLM with the ability to sequentially acquire new alignment preferences without eroding past judgments, we introduce **Focalized Preference Optimization (FPO)**, which focuses learning where it is needed and eases off where mastery has already been achieved.

Our approach builds upon the theoretical foundation of Direct Preference Optimization (DPO) (Rafailov et al. 2023), which aims to maximize the margin between a preferred response y^p (chosen/preferred) and a dispreferred one y^d (rejected/dispreferred). This margin can be expressed by

the implicit reward r:

$$r = \beta \left(\log \frac{\pi_\theta(y^p \mid x)}{\pi_{\text{ref}}(y^p \mid x)} - \log \frac{\pi_\theta(y^d \mid x)}{\pi_{\text{ref}}(y^d \mid x)} \right), \quad (1)$$

where $\pi_\theta(y \mid x)$ denotes the probability assigned by the fine-tuned model with parameters θ, $\pi_{\text{ref}}(y \mid x)$ is the corresponding probability under the unadapted reference model, and β controls the sharpness of preference. The standard DPO loss is calculated as:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma(r). \quad (2)$$

While effective for static learning, this formulation treats every sample with equal force. Even with a rehearsal mechanism, repeatedly pressing on well-learned pairs can perpetuate a drift that overrides earlier tasks. Our FPO reframes preference optimization as adaptive attention to uncertainty. We define the FPO loss as:

$$\mathcal{L}_{\text{FPO}} = -(1 - \sigma(r))^2 \log \sigma(r), \quad (3)$$

where $(1 - \sigma(r))^2$ functions as a gating term that scales the gradient according to the model's confidence in its current preference alignment.

This gating term creates a dual-regime learning behavior. For new or poorly-studied samples (where $r \leq 0$), $\sigma(r)$ is relatively small (close to 0.5 or less), and consequently, $(1 - \sigma(r))^2$ remains close to 1. In this scenario, $\mathcal{L}_{\text{FPO}}$ is nearly identical to $\mathcal{L}_{\text{DPO}}$, allowing the optimization to deliver a full corrective signal. This enables the model to effectively learn new preferences or review those that have been forgotten. Conversely, for well-studied historical samples (where $r > 0$), both $\sigma(r)$ and $(1 - \sigma(r))^2$ lie strictly between 0 and 1. As the reward gap between chosen and rejected responses grows (r increases), $\sigma(r)$ becomes larger, and the gating term $(1 - \sigma(r))^2$ shrinks. The gradient is not eliminated but is significantly attenuated in proportion to the model's increasing confidence. This allows for fine-grained adjustments without exerting unnecessary pressure on already aligned pairs, thereby allocating most of the learning capacity to uncertain examples while preserving previously acquired preferences.

To strengthen selective adaptation within the lifelong learning framework, we introduce a rehearsal mechanism based on a fixed-size buffer. For each new task, the buffer is merged with current data, shuffled, and trained jointly. After training, 20% of the new task data are randomly sampled and added to the buffer, removing the oldest samples if the buffer is full (Ahn and Kim 2025). This replay of prior data, together with a focalized loss, preserves established knowledge while acquiring new ones, akin to human rehearsal that reinforces key memories without redundancy. The dual strategy in FPO mitigates catastrophic forgetting and value drift, ensuring new learning builds upon prior alignment.

Short-to-Long Memory Consolidation

Although Focalized Preference Optimization (FPO) offers targeted learning, unchecked parameter updates can still cause knowledge interference and catastrophic forgetting over time. Human cognition solves this through memory consolidation, a process where fresh, labile memories are gradually stabilized and integrated into the neocortex for long-term storage (Tulving and Thomson 1973; Baddeley 2000). Inspired by this, we introduce the **Short-to-Long Memory Consolidation (SLMC)** module. SLMC transforms the raw, ephemeral parameter changes from a single task's FPO training into a durable, refined update that harmonizes with the model's accumulated wisdom, ensuring continuous and stable alignment.

The SLMC process unfolds in three stages, as depicted in Figure 2. Before processing task t, the model's accumulated alignment knowledge is implicitly stored within its current long-term memory, represented by the model parameters LM_{t-1}. Upon completing FPO training on task t, we denote the resulting LoRA parameter update as the raw short-term memory trace, SM_t. Since SM_t may contain both high-frequency noise and components that conflict with previously consolidated knowledge, SLMC refines SM_t into a stable, conflict-aware trace RSM_t, which is then merged back into the long-term memory.

Denoising Short-Term Memory. The initial SM_t captures not only the essential alignment preference of the new task but also ephemeral, noisy artifacts from the training process. To distill the core knowledge, we treat each LoRA matrix update ($\Delta W_{A/B} \in \mathbb{R}^{d \times r}$, where d is the hidden dimension and r is the LoRA rank) within SM_t as a signal to be purified. Inspired by the Eckart–Young–Mirsky theorem (Eckart and Young 1936; Mirsky 1960), we firstly perform Singular Value Decomposition (SVD) on the short-term memory matrix SM_t:

$$\text{SM}_t = U_t \Sigma_t V_t^T, \quad (4)$$

where $U_t \in \mathbb{R}^{d \times k}$, $\Sigma_t \in \mathbb{R}^{k \times k}$ is a diagonal matrix of singular values, and $V_t \in \mathbb{R}^{r \times k}$, with $k \leq \min(d, r)$ being the effective rank corresponding to the number of non-zero singular values. The diagonal entries of Σ_t quantify the energy of each singular direction: the largest singular values capture foundational alignment modifications, whereas the smaller values predominantly reflect noise.

We then apply a denoising procedure by preserving only the most significant components. With the singular values $\{\sigma_i\}_{i=1}^k$ from Σ_t sorted in descending order, we seek the smallest rank k' that captures at least a fraction θ (e.g., 0.9) of the total signal energy. This is determined by finding the minimum k' that satisfies:

$$\frac{\sum_{i=1}^{k'} \sigma_i^2}{\sum_{i=1}^{k} \sigma_i^2} \geq \theta. \quad (5)$$

We then truncate each SVD matrix to its leading k' components, formulated as:

$$U_t' = U_t[:, :k'], \Sigma_t' = \Sigma_t[:k', :k'], V_t' = V_t[:, :k']. \quad (6)$$

The denoised update SM_t' is then reconstructed as:

$$\text{SM}_t' = U_t' \Sigma_t' V_t'^T. \quad (7)$$

This acts as a low-rank filter, retaining the principal directions of change while discarding high-frequency, task-specific noise, thereby yielding a more generalized and robust short-term memory trace.

Conflict-Aware Refinement. Next, we must ensure that the new, denoised knowledge SM'_t does not destructively overwrite critical past learnings. We achieve this by projecting SM'_t onto a knowledge subspace spanned by historical refined updates. We first flatten the LoRA update matrices into vectors. Let $n = d \times r$ be the dimensionality of the flattened LoRA parameter vector. We then stack the flattened refined short-term memory vectors from previous tasks $\{\mathrm{RSM}_j\}_{j=1}^{t-1}$ into a matrix H:

$$H = \begin{bmatrix} \mathrm{RSM}_1 \\ \mathrm{RSM}_2 \\ \vdots \\ \mathrm{RSM}_{t-1} \end{bmatrix} \in \mathbb{R}^{(t-1)\times n}. \tag{8}$$

To extract the main axes representing consistent patterns in past updates, we compute the economy SVD of the matrix $H = U_h \, \Sigma_h \, V_h^T$. The columns of $V_h \in \mathbb{R}^{n \times k_h}$ form an orthonormal basis $\{v_j\}_{j=1}^{k_h}$ for the knowledge subspace, where k_h is the rank of matrix Σ_h. The projection of the flattened denoised vector SM'_t onto this subspace is calculated as:

$$\mathrm{SM}_t^{\mathrm{p}} = V_h \, V_h^T \, \mathrm{SM}'_t = \sum_{j=1}^{k_h} \langle v_j, \, \mathrm{SM}'_t \rangle \, v_j, \tag{9}$$

which captures the part of SM'_t that conflicts with the historical alignment signals. Correspondingly, the orthogonal component $\mathrm{SM}_t^{\mathrm{o}} = \mathrm{SM}'_t - \mathrm{SM}_t^{\mathrm{p}}$ represents the truly novel and safe information that is orthogonal to past knowledge.

To mitigate interference with established memories, we selectively suppress the projected component by scaling it with a hyperparameter scaling factor $\lambda \in [0, 1]$, while preserving the novel orthogonal component intact:

$$\mathrm{RSM}_t = \mathrm{SM}_t^{\mathrm{o}} + \lambda \, \mathrm{SM}_t^{\mathrm{p}}, \tag{10}$$

where $\mathrm{RSM}_t \in \mathbb{R}^n$ is the final refined short-term memory vector. This step allows the model to learn without overwriting old memories, mirroring how the hippocampus preserves cortical memories.

Long-Term Memory Integration Finally, the refined, conflict-free update RSM_t is ready for permanent integration. After being reshaped from a vector in $\mathbb{R}^n$ back to a matrix in $\mathbb{R}^{d \times r}$, the update is directly added to the pre-task parameters to form the new long-term memory as follows:

$$\mathrm{LM}_t = \mathrm{LM}_{t-1} + \mathrm{reshape}(\mathrm{RSM}_t). \tag{11}$$

The consolidated state LM_t now serves as the stable foundation for the next round of lifelong learning. Crucially, the refined update RSM_t is also added to our historical memory bank H to inform future conflict resolution. Through this cycle of distillation, conflict resolution, and integration, SLMC enables the LLM to continuously evolve its alignment preferences and values, ensuring robustness against catastrophic forgetting in lifelong learning scenarios.

4 Experimental Results

Experiment Setup

Datasets. We introduce a comprehensive benchmark for lifelong alignment from six diverse datasets to evaluate four key dimensions. It assesses: (I) Human Preference Alignment (HPA) using HC3 (Guo et al. 2023) and hh-rlhf-helpful (Bai et al. 2022a); (ii) Instruction Fidelity Alignment (IFA) with Capybara-Preferences (Argilla 2024); (iii) Value Alignment (VA) with hh-rlhf-harmless (Bai et al. 2022a) and Safe-RLHF (Dai et al. 2023); and (iv) Objective Factual Alignment (OFA) with TruthfulQA (Lin, Hilton, and Evans 2022). Existing benchmarks suffer from significant limitations in scope and coverage. For instance, CPPO relies exclusively on the Reddit TL;DR dataset, focusing narrowly on IFA without addressing broader alignment concerns (Zhang et al. 2024). While COPR offers more breadth by incorporating datasets for HPA, IFA, and VA, it omits any evaluation of objective factual consistency (Zhang et al. 2025). In contrast, our six-dataset suite provides a more holistic and rigorous evaluation framework, spanning the full spectrum, from subjective value judgments to objective truthfulness.

Evaluation Metrics. Following (Chaudhry et al. 2018; Lopez-Paz and Ranzato 2017), we evaluate lifelong alignment performance using three standard metrics to assess retention, interference, and overall effectiveness, including Last Performance (Last), Backward Transfer (BWT), and Average Performance (AP). The performance on task j after training to task i, denoted as $m_{i,j}$, is calculated using BLEU-4 (Papineni et al. 2002), ROUGE-L (Lin 2004), and LLM-Judge score. Our LLM-Judge utilizes the DeepSeek-Chat API, along with six self-designed, task-specific prompt templates, to evaluate response quality, with further details available in the supplementary materials. Three metrics are computed as follows: $\mathrm{Last} = \frac{1}{N}\sum_{i=1}^{N} m_{N,i}$, $\mathrm{BWT} = \frac{1}{N-1}\sum_{i=1}^{N-1}(m_{N,i} - m_{i,i})$, $\mathrm{AP} = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{k}\sum_{i=1}^{k} m_{k,i}$ where N is the number of total tasks.

Baselines. We evaluate several representative strategies: vanilla sequence finetuning (SeqFT), replay-based methods like ER (Rolnick et al. 2019) and GEM (Lopez-Paz and Ranzato 2017); regularization-based approaches like EWC (Chaudhry et al. 2018) and O-LoRA (Wang et al. 2023a); and the architecture-based method like L2P (Wang et al. 2022). In addition, we re-implement CPPO (Zhang et al. 2024) to evaluate its performance in the lifelong alignment setting. We also include single-task learning (STL) and multi-task learning (MTL) as strong upper bounds to contextualize the performance of the lifelong learning approaches. To enable lifelong alignment, both the supervised fine-tuning (SFT) initialization and direct preference optimization (DPO) phases for each baseline method.

Implementation Details. We train our models on eight A800-80GB GPUs using the LLaMA-Factory framework [1]. The backbone model is Qwen-2.5-7b-Instruct. We perform Supervised Fine-Tuning (SFT) for 3 epochs with a learning rate of 1e-4, followed by Direct Preference Optimization (DPO) for 3 epochs with a learning rate of 5e-6. Our training datasets are ordered as Task 1 to Task 6, corresponding to Capybara-Preferences, HC3, hh-rlhf-harmless-base, hh-rlhf-helpful-base, Safe-RLHF, and TruthfulQA, respec-

[1] https://github.com/hiyouga/LLaMA-Factory

Methods	BLEU-4			ROUGE-L			LLM-Judge			AVG		
	BWT	Last	AP	BWT	Last	AP	BWT	Last	AP	BWT	Last	AP
SeqFT	-19.06	11.53	18.34	-11.82	15.76	18.41	-9.21	39.72	39.42	-13.36	22.34	25.39
L2P	**0.67**	12.59	17.61	**1.61**	9.80	11.81	**3.26**	34.04	34.03	**1.85**	18.81	21.15
O-LoRA	-11.16	15.01	25.85	-1.93	6.83	17.82	-18.28	48.66	**58.24**	-10.46	23.50	33.98
ER	-6.29	<u>22.73</u>	<u>30.39</u>	-0.42	<u>24.81</u>	<u>24.46</u>	-6.66	<u>49.11</u>	52.03	-4.46	<u>32.22</u>	<u>35.63</u>
GEM	-16.14	14.53	19.63	-9.10	18.11	19.48	-10.47	45.15	44.20	-11.90	25.93	27.77
EWC	-10.22	15.09	25.76	-4.21	17.04	21.46	-11.82	46.15	50.89	-8.75	26.09	32.70
CPPO	-10.85	13.50	19.96	-4.15	18.12	18.78	-10.2	46.05	45.64	-8.40	25.89	28.13
LifeAlign (Ours)	<u>0.02</u>	**29.14**	**30.53**	<u>1.39</u>	**26.43**	**24.84**	<u>1.31</u>	**57.42**	<u>53.91</u>	<u>0.91</u>	**37.67**	**36.43**
STL	–	28.72	–	–	27.27	–	–	54.84	–	–	36.94	–
MTL (Upper Bound)	–	30.64	–	–	26.34	–	–	57.01	–	–	38.00	–

Table 1: Performance (%) of our method and distinct lifelong learning methods. The best and suboptimal results are emphasized in **bold** and <u>underline</u>. The last three columns represent the average values of the three metrics.

tively. We also vary this sequence to evaluate the robustness of our method to different task orderings. Detailed experimental settings are in the supplementary material.

Main Results

Table 1 presents the main results of our proposed method, LifeAlign, against various lifelong learning baselines. **First**, LifeAlign achieves state-of-the-art performance across almost all metrics, with scores rivaling the MTL upper bound. Unlike replay-based approaches such as ER, which cause knowledge interference (AVG Last 32.22), or overly conservative regularization methods like EWC (only 32.70 on AVG AP), LifeAlign uses its SLMC module to distill and integrate knowledge non-destructively. This conflict-aware consolidation, building upon FPO's targeted learning signal, allows the model to develop a coherent value system and attain superior performance. **Second**, LifeAlign demonstrates exceptional resistance to catastrophic forgetting, maintaining a positive BWT in stark contrast to baselines like O-LoRA (-18.28 BWT) and GEM (-10.47 BWT), which suffer severe degradation. At the training level, FPO's adaptive loss preserves existing knowledge from being overwritten. Subsequently, at the parameter level, SLMC resolves destructive conflicts before integration, ensuring robust knowledge preservation. **Third**, LifeAlign outperforms two specific baselines: L2P and CPPO. L2P's prompt-based isolation mitigates forgetting (positive AVG BWT of 1.85) but sacrifices performance, leading to very low Last (18.81) and AP (21.15). CPPO, while making a sophisticated effort to balance sample contributions, just considers a sample's immediate impact on current parameters, resulting in severe forgetting (AVG BWT -8.40). In contrast, LifeAlign's dual-component design overcomes these limitations by using FPO to manage sample-level plasticity and SLMC to ensure structural, parameter-level stability, thereby achieving a more robust and effective balance.

Ablation Results

To evaluate the individual contributions of our core components, we conduct an ablation study by systematically removing FPO and the SLMC module, with results in Table 2.

	Modules		BLEU-4			LLM–Judge		
	FPO	SLMC	BWT	Last	AP	BWT	Last	AP
a	✗	✗	-7.67	21.25	28.95	-7.02	49.11	51.81
b	✗	✓	-1.39	27.12	29.97	-0.01	55.63	52.28
c	✓	✗	-3.68	25.40	30.05	-4.47	51.83	52.09
d	✓	✓	0.02	29.14	30.53	1.31	57.42	53.91

Table 2: The results of the ablation study.

The experimental results demonstrate the critical and complementary roles of both components. The baseline model (row a), without either, suffers from severe catastrophic forgetting, shown by its highly negative BWT (-7.67 BLEU-4 and -7.02 LLM-Judge). Introducing only the SLMC module (row b) yields a significant improvement in mitigating forgetting, dramatically increasing the BWT to -1.39 and -0.01 respectively, highlighting its effectiveness in resolving conflicts between task updates and preserving historical knowledge. Conversely, using only FPO (row c) reduces forgetting to a lesser extent, indicating its adaptive loss is helpful but insufficient on its own to prevent knowledge erosion. Ultimately, the full LifeAlign method (row d), which integrates both components, achieves the best performance across all metrics and is the only configuration to produce a positive BWT. This synergistic result validates our design: FPO provides a more targeted and stable learning signal during alignment, while SLMC then effectively distills and integrates it into the model's long-term memory, leading to robust and continuous alignment.

Further Analysis

Impact of Hyperparameters. We perform a sensitivity analysis on the two key hyperparameters of our SLMC module: the denoising threshold θ and the projection weight λ. First, with θ fixed at 0.9, we vary λ from 0 to 1. As shown in Figure 3 (a) and (b), performance peaks at $\lambda = 0.5$, indicating an optimal balance between retaining historical knowledge and incorporating new, conflict-free updates. Values of λ that are too low fail to preserve sufficient prior knowledge,

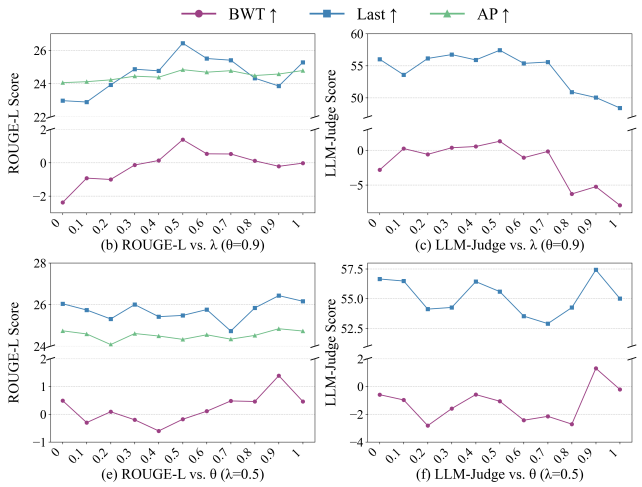

Figure 3: Performance sensitivity of hyperparameters.

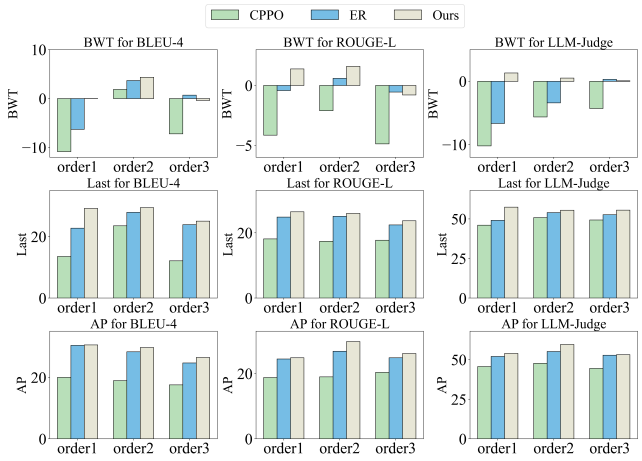

Figure 4: Results of different task order.

while those that are too high overly constrain new learning, harming performance. Next, fixing $\lambda = 0.5$, we evaluate θ from 0 to 1. Results in Figure 3 (c) and (d) show a clear peak at $\theta = 0.9$: performance improves with higher thresholds but declines beyond this point. This suggests that preserving 90% of the information effectively captures core alignment signals while filtering out high-frequency noise. Lower thresholds risk removing useful information, while higher ones risk reintroducing noise. Based on this analysis, we set $\theta = 0.9$ and $\lambda = 0.5$ for all main experiments.

Impact of Task Order. Figure 4 evaluates the impact of task order on LifeAlign and two representative baselines (ER and CPPO), using three sequences: forward ($1\rightarrow2\rightarrow\ldots\rightarrow6$), reverse ($6\rightarrow5\rightarrow\ldots\rightarrow1$), and random ($3\rightarrow1\rightarrow6\rightarrow4\rightarrow2\rightarrow5$). LifeAlign consistently outperforms both baselines across all orders. For example, on the LLM-Judge metric under the reverse sequence, LifeAlign achieves a positive LLM-Judge BWT of 0.51, while CPPO and ER suffer severe forgetting with BWTs of -5.64 and -3.41. Notably, baseline performance varies significantly with task order, whereas LifeAlign remains stable. Its Last score stays high, and BWT remains positive or near zero across all sequences, indicating minimal forgetting. In contrast, CPPO shows high variance and consistent catastrophic forgetting regardless of order. These results demonstrate that LifeAlign is robust to task ordering, a crucial advantage for real-world lifelong alignment systems, as the arrival of new preferences and norms is typically unpredictable.

Backbone	Methods	BLEU-4			ROUGE-L			LLM–Judge		
		BWT	Last	AP	BWT	Last	AP	BWT	Last	AP
Qwen	ER	-6.29	22.73	30.39	-0.42	24.81	24.46	-6.66	49.11	52.03
	CPPO	-10.85	13.50	19.96	-4.15	18.12	18.78	-10.2	46.05	45.64
	Ours	0.02	29.14	30.53	1.39	26.43	24.84	1.31	57.42	53.91
Mistral	ER	-0.69	31.50	31.46	-2.86	25.22	25.4	-3.51	59.42	60.48
	CPPO	-9.86	12.85	16.88	-4.83	17.16	16.64	0.04	47.05	44.92
	Ours	1.76	32.39	31.50	-1.10	27.82	25.84	-1.74	60.08	61.71
LLaMA	ER	0.57	29.70	30.94	0.13	27.18	25.47	-2.62	57.31	60.93
	CPPO	-13.80	11.45	16.18	-6.82	15.26	15.40	-7.77	47.57	46.07
	Ours	1.86	30.40	31.42	1.62	28.48	25.88	2.10	60.84	62.58

Table 3: Performance(%) of foundation models across three methods with the default order.

Influence of Foundation Models. To evaluate the generalizability of our framework, we test LifeAlign on three distinct foundation models: Mistral-7B-v0.3, Qwen-2.5-7B-Instruct, and LLaMA-3.1-8B-Instruct. Due to space constraints, we compare against two representative baselines (ER and CPPO) in Table 3. LifeAlign consistently outperforms both baselines across all architectures, achieving the highest Last and AP, especially on the holistic LLM-Judge metric. Notably, LifeAlign effectively mitigates catastrophic forgetting across all backbones, maintaining positive or near-zero BWT. In contrast, CPPO exhibits severe forgetting, with BWT dropping to -7.77 on LLaMA, while LifeAlign achieves a favorable 2.1. These results demonstrate that LifeAlign is robust and broadly effective, regardless of the underlying model architecture.

5 Conclusions and Future Work

In this paper, we introduced LifeAlign, a novel framework that addresses catastrophic forgetting in lifelong LLM alignment. Our approach enables models to sequentially adapt to new preferences by integrating two innovations: Focalized Preference Optimization (FPO), which intelligently directs learning towards new or uncertain preferences while preserving established ones, and Short-to-Long Memory Consolidation (SLMC), a cognitively-inspired mechanism that distills, refines, and stably integrates new knowledge. Comprehensive experiments demonstrate that LifeAlign significantly outperforms existing strategies in both knowledge retention and final alignment quality across diverse tasks and models. Future directions include enhancing the scalability and computational efficiency of LifeAlign, developing a rehearsal-free variant to address growing privacy concerns, and ultimately deploying it in live, interactive systems for real-world validation.

Acknowledgments

The authors wish to thank the reviewers for their helpful comments and suggestions. This research is funded by the National Nature Science Foundation of China (No. 62477010, No.62577022, No.62307028 and No.62477012), the Natural Science Foundation of Shanghai (No. 23ZR1441800 and No.23ZR1418500), Shanghai Science and Technology Innovation Action Plan (No. 24YF2710100 and No.23YF1426100), Shanghai Qiji Zhifeng Co., Ltd. (2025-GZL-RGZN-01001) and the opening funding of the State Key Laboratory of DisasterReduction in Civil Engineering (Grant No. SLDRCE24-03).

References

Ahn, S.; and Kim, S. B. 2025. Impact of Replay Ratios on Performance and Efficiency in Continual Learning for Skeleton-Based Action Recognition. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 460–466. Springer.

Argilla. 2024. Capybara-Preferences Dataset. https://huggingface.co/datasets/argilla/Capybara-Preferences.

Baddeley, A. 2000. The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, 4(11): 417–423.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Chaudhry, A.; Dokania, P. K.; Ajanthan, T.; and Torr, P. H. 2018. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, 532–547.

Dai, J.; Pan, X.; Sun, R.; Ji, J.; Xu, X.; Liu, M.; Wang, Y.; and Yang, Y. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Eckart, C.; and Young, G. 1936. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3): 211–218.

Guo, B.; Zhang, X.; Wang, Z.; Jiang, M.; Nie, J.; Ding, Y.; Yue, J.; and Wu, Y. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Huai, T.; Zhou, J.; Cai, Y.; Chen, Q.; Wu, W.; Wu, X.; Qiu, X.; and He, L. 2025. Task-Core Memory Management and Consolidation for Long-term Continual Learning. arXiv:2505.09952.

Jang, J.; Ye, S.; Lee, C.; Yang, S.; Shin, J.; Han, J.; Kim, G.; and Seo, M. 2022. Temporalwiki: A lifelong benchmark for training and evaluating ever-evolving language models. *arXiv preprint arXiv:2204.14211*.

Jin, Q.; Yang, Y.; Chen, Q.; and Lu, Z. 2023. Genegpt: augmenting large language models with domain tools for improved access to biomedical information. arXiv. *Ovadia, O., Brief, M., Mishaeli, M., & Elisha, O.(2023). Fine-tuning or retrieval*.

Ke, Z.; Shao, Y.; Lin, H.; Konishi, T.; Kim, G.; and Liu, B. 2023. Continual pre-training of language models. *arXiv preprint arXiv:2302.03241*.

Lee, H.; Phatale, S.; Mansoor, H.; Mesnard, T.; Ferret, J.; Lu, K. R.; Bishop, C.; Hall, E.; Carbune, V.; Rastogi, A.; et al. 2024. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. In *International Conference on Machine Learning*, 26874–26901. PMLR.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252.

Lopez-Paz, D.; and Ranzato, M. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30.

Lu, X.; Yu, B.; Lu, Y.; Lin, H.; Yu, H.; Sun, L.; Han, X.; and Li, Y. 2024. SoFA: Shielded On-the-fly Alignment via Priority Rule Following. In *Findings of the Association for Computational Linguistics ACL 2024*, 7108–7136.

Mirsky, L. 1960. Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics*, 11(1): 50–59.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Que, H.; Liu, J.; Zhang, G.; Zhang, C.; Qu, X.; Ma, Y.; Duan, F.; Bai, Z.; Wang, J.; Zhang, Y.; et al. 2024. D-cpt law: Domain-specific continual pre-training scaling law for large language models. *Advances in Neural Information Processing Systems*, 37: 90318–90354.

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. *Advances in neural information processing systems*, 36: 53728–53741.

Rolnick, D.; Ahuja, A.; Schwarz, J.; Lillicrap, T.; and Wayne, G. 2019. Experience replay for continual learning. *Advances in neural information processing systems*, 32.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shi, H.; Xu, Z.; Wang, H.; Qin, W.; Wang, W.; Wang, Y.; Wang, Z.; Ebrahimi, S.; and Wang, H. 2024. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*.

Tulving, E.; and Thomson, D. M. 1973. Encoding specificity and retrieval processes in episodic memory. *Psychological review*, 80(5): 352.

Wang, X.; Chen, T.; Ge, Q.; Xia, H.; Bao, R.; Zheng, R.; Zhang, Q.; Gui, T.; and Huang, X. 2023a. Orthogonal Subspace Learning for Language Model Continual Learning. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10658–10671. Singapore: Association for Computational Linguistics.

Wang, X.; Zhang, Y.; Chen, T.; Gao, S.; Jin, S.; Yang, X.; Xi, Z.; Zheng, R.; Zou, Y.; Gui, T.; et al. 2023b. Trace: A comprehensive benchmark for continual learning in large language models. *arXiv preprint arXiv:2310.06762*.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 139–149.

Wu, T.; Luo, L.; Li, Y.-F.; Pan, S.; Vu, T.-T.; and Haffari, G. 2024. Continual Learning for Large Language Models: A Survey. arXiv:2402.01364.

Xie, Y.; Aggarwal, K.; and Ahmad, A. 2024. Efficient continual pre-training for building domain specific large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, 10184–10201.

Xu, C.; Chern, S.; Chern, E.; Zhang, G.; Wang, Z.; Liu, R.; Li, J.; Fu, J.; and Liu, P. 2023. Align on the fly: Adapting chatbot behavior to established norms. *arXiv preprint arXiv:2312.15907*.

Yadav, P.; Sun, Q.; Ding, H.; Li, X.; Zhang, D.; Tan, M.; Bhatia, P.; Ma, X.; Nallapati, R.; Ramanathan, M. K.; et al. 2023. Exploring Continual Learning for Code Generation Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 782–792.

Yang, Y.; Zhou, J.; Ding, X.; Huai, T.; Liu, S.; Chen, Q.; Xie, Y.; and He, L. 2025a. Recent advances of foundation language models-based continual learning: A survey. *ACM Computing Surveys*, 57(5): 1–38.

Yang, Y.; Zhou, J.; Li, J.; Pan, Q.; Zhan, B.; Chen, Q.; Qiu, X.; and He, L. 2025b. Reinforced Interactive Continual Learning via Real-time Noisy Human Feedback. arXiv:2505.09925.

Zhang, H.; Gui, L.; Lei, Y.; Zhai, Y.; Zhang, Y.; Zhang, Z.; He, Y.; Wang, H.; Yu, Y.; Wong, K.-F.; Liang, B.; and Xu, R. 2025. COPR: Continual Human Preference Learning via Optimal Policy Regularization. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 5377–5398. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Zhang, H.; Lei, Y.; Gui, L.; Yang, M.; He, Y.; WANG, H.; and Xu, R. 2024. CPPO: Continual Learning for Reinforcement Learning with Human Feedback. In Kim, B.; Yue, Y.; Chaudhuri, S.; Fragkiadaki, K.; Khan, M.; and Sun, Y., eds., *International Conference on Representation Learning*, volume 2024, 22719–22742.

Zhu, M.; Liu, Y.; Zhang, L.; Guo, J.; and Mao, Z. 2025. On-the-fly preference alignment via principle-guided decoding. *arXiv preprint arXiv:2502.14204*.