

# Parameter Importance-Driven Continual Learning for Foundation Models

Lingxiang Wang<sup>1,2</sup> Hainan Zhang<sup>1,2\*</sup> Zhiming Zheng<sup>1,2</sup>

<sup>1</sup>Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, Beihang University

<sup>2</sup>School of Artificial Intelligence, Beihang University

wanglingxiang@buaa.edu.cn, zhanghainan@buaa.edu.cn, zhengzhiming0130@163.com

## Abstract

Domain-specific post-training often causes catastrophic forgetting, making foundation models lose their general reasoning ability and limiting their adaptability to dynamic real-world environments. Preserving general capabilities while acquiring downstream domain knowledge is a central challenge for large language and multimodal models. Traditional continual learning methods, such as regularization, replay and architectural isolation, suffer from poor downstream performance, reliance on inaccessible historical data, or additional parameter overhead. While recent parameter-efficient tuning (PET) methods can alleviate forgetting, their effectiveness strongly depends on the choice of parameters and update strategies. In this paper, we introduce PIECE, a Parameter Importance Estimation-based Continual Enhancement method that preserves general ability while efficiently learning domain knowledge without accessing prior training data or increasing model parameters. PIECE selectively updates only 0.1% of core parameters most relevant to new tasks, guided by two importance estimators: PIECE-F based on Fisher Information, and PIECE-S based on a second-order normalization that combines gradient and curvature information. Experiments across three language models and two multimodal models show that PIECE maintains general capabilities and achieves state-of-the-art continual learning performance across diverse downstream tasks. Our results highlight a practical path to scalable, domain-adaptive foundation models without catastrophic forgetting.

## 1. Introduction

Foundation models, including large language models (LLMs) and multimodal LLMs (MLLMs), have shown remarkable generalization and reasoning abilities across diverse domains, from natural language understanding [23,

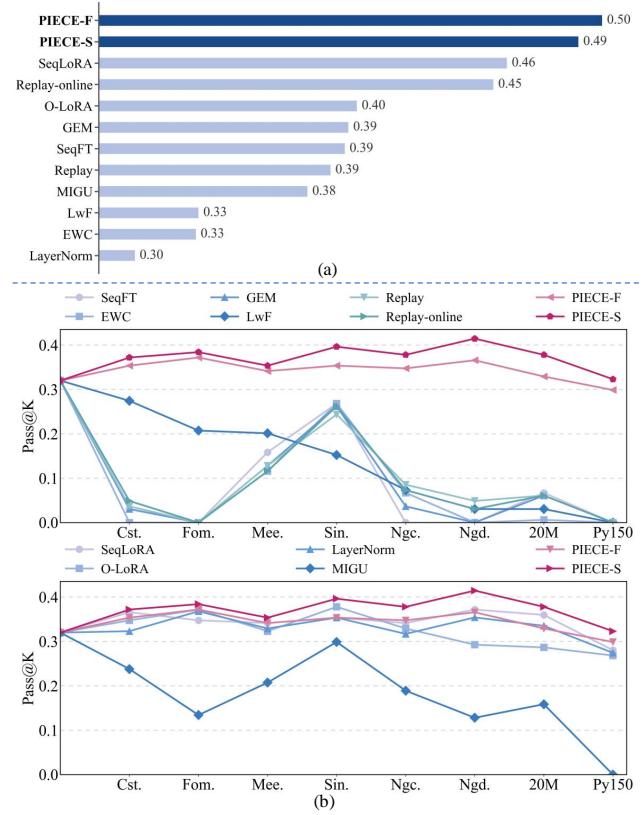


Figure 1. (a) Average downstream-task scores and (b) HumanEval Programming ability (Pass@K, K=1) of Llama3-8B on the TRACE benchmark as continual learning tasks increase. PIECE consistently outperforms full fine-tuning (SeqFT), regularization (EWC, GEM, LwF), replay (Replay, Replay-online), and PET (SeqLoRA, O-LoRA, LayerNorm, MIGU) baselines in both downstream performance and capability retention.

[69] and multimodal reasoning [28, 42] to applications in healthcare [11] and science [59]. However, when fine-tuned on new domains, these models often suffer from catastrophic forgetting [17, 31], which erodes their core strengths, such as general reasoning abilities and broad knowledge, and ultimately limiting their effectiveness in

\*Corresponding author

domain-specific adaptation [30, 37].

Continual learning [40, 47] has been recognized as an effective mechanism for enabling models to incrementally acquire new domain knowledge. **Regularization methods** [24, 34] constrain model updates to remain close to the original parameters, but it may hinder efficient learning for downstream tasks. **Replay methods** [44, 45] rely on historical training data, which is impractical for post-training foundation models [54] and inefficient due to large-scale data replay. **Architectural isolation methods** [53, 67] mitigate forgetting by separating different tasks through mechanisms such as gating, but they introduce additional parameter overhead and impede cross-task generalization [55]. To compare these methods, we take programming ability as a representative proxy for the general capabilities of foundation models, because it is particularly vulnerable to destruction [8]. Figure 1 (b) evaluate the forgetting behaviors of these traditional approaches, showing that both Regularization and Replay methods<sup>1</sup> fail to meet the crucial requirement of preserving the broad and general capabilities of foundation models during continual learning.

Recently, parameter-efficient tuning (PET) methods [14, 21, 56, 65] have demonstrated the potential for anti-forgetting learning [50, 58] without relying on historical training data, task labels, or modifying model architecture. As shown in Figure 1, our analysis reveals that: **(1) PET outperforms traditional continual learning approaches in preserving the original capabilities of the model.** In contrast, regularization and replay methods struggle with catastrophic forgetting due to the lack of access to historical training data. **(2) Different PET strategies exhibit varying levels of general capability preservation.** SeqLoRA [21], which updates only Attention components, outperforms LayerNorm [65] updates on downstream performance, suggesting that domain knowledge is possibly concentrated in Attention components [29]. However, it slightly weakens original abilities due to all layers structural disruption (see Section 4.2). Thus, the choice of parameters to fine-tune and the update strategy play critical roles in domain transfer and prior knowledge retention.

In this paper, we propose PIECE, a Parameter Importance Estimation-based Continual Enhancement method that effectively learns domain-specific knowledge while preserving general capabilities. **PIECE focuses on the internal model updates without the need for historical data or additional parameters.** It selectively updates only the most relevant 0.1% of core parameters for new tasks. Specifically, before fine-tuning, PIECE evaluates parameter importance using two independent strategies: one based on Fisher information (PIECE-F), which measures parameter sensitivity to the current task loss, and the other based

<sup>1</sup>We only focus on continue learning without parameter growth, thus we don't compare the architectural isolation methods.

on a second-order normalization method (PIECE-S), which combines gradient and curvature information to assess parameter importance. Notably, PIECE-S is our first theory-grounded variant and better preserves original model capabilities than PIECE-F. Based on these importance scores, only the Top-0.1% parameters are updated during finetuning, while the remaining parameters are frozen. This selective updating enables efficient domain adaptation while preserving the model's original capability.

We conducted extensive experiments<sup>2</sup> on three LLMs and two MLLMs, encompassing model sizes ranging from 2B to 14B. The results demonstrate that PIECE effectively preserves prior knowledge across diverse downstream tasks while achieving state-of-the-art continual learning performance, surpassing strong baselines. Our analysis reveals that PIECE selectively updates portions of Attention as well as LayerNorm modules at lower and deeper layers, enabling it to maintain general capabilities while efficiently assimilating new knowledge. These findings highlight a practical and scalable pathway toward domain-adaptive foundation models that mitigate catastrophic forgetting. Our contributions can be summarized as follows:

- We find that parameter selection is an effective strategy to mitigate catastrophic forgetting, enabling continual learning through internal model updates without the need for historical data or additional parameters.
- We propose two parameter importance estimation methods: PIECE-F and PIECE-S. Notably, PIECE-S is our first theory-grounded variant and better preserves original model capabilities than PIECE-F.
- PIECE is a model-agnostic, scalable method that enables sustainable continual learning while preserving pre-trained abilities, as validated by extensive experiments across diverse LLMs and MLLMs.

## 2. Related Work

**Regularization-based methods** mitigate catastrophic forgetting by constraining changes to weights or units associated with previous tasks. Some research focuses on preserving task-relevant parameters [3, 24], while others emphasize constraints on the model's intermediate or final outputs [27]. **Replay-based methods** introduce the fixed memory to store either real samples [45, 70] or pseudo-generative examples [38, 48] from previous tasks, and replay them during training to preserve the acquired knowledge. Although these two types of methods have shown significant effectiveness in mitigating catastrophic forgetting, access to pretraining data is often limited in foundation models for ordinary users, which constrains their practical applicability. **Architecture-based methods** ex-

<sup>2</sup>See code, dataset, logs and appendix in Supplemental Materials and <https://github.com/wanglingxiang0717/PIECE>.

pand or isolate model parameters to retain new knowledge while preventing the forgetting of prior knowledge. Inspired by Aljundi et al. [2], some methods [10, 25] employ gating mechanisms to assign independent modules to different tasks, thereby achieving knowledge isolation. Tong et al. [53] further optimize the training of the gates to improve classification accuracy and overall model performance, while Zhao et al. [67] enhance the design of the gates to increase the utilization efficiency of task-specific modules, improving knowledge transfer. However, these methods cause the parameter size to grow linearly with the number of tasks, resulting in higher computational and storage costs. **Parameter-Efficient Tuning** (PET) [13, 62] has become an active research direction in post-training of large models. Some research has shown that parameter-efficient tuning offers significant advantages in mitigating catastrophic forgetting [5, 6], with LoRA [21] demonstrating particularly strong performance [67]. Wang et al. [56] proposed O-LoRA, a method that updates LoRA weights within an orthogonal subspace to enhance the model’s continual learning capability. Compared to these structured methods, unstructured parameter-efficient tuning focuses more on the model’s transfer performance on downstream tasks. Sung et al. [50] and Xu et al. [58] demonstrated that tuning only a subset of task-relevant parameters can achieve better transfer performance than full fine-tuning. Subsequently, Du et al. [14] introduce MIGU, which uses unstructured parameter-efficient tuning for continual learning to retain model capabilities without historical data. However, its parameter importance relies solely on the L1-norm of outputs. Building on this, we explore unstructured parameter-efficient tuning and propose two improved parameter importance evaluation strategy.

### 3. Method

In this section, we introduce the components of PIECE, including preliminaries, two types of parameter importance estimators (PIECE-F and PIECE-S) and mask-based implementations. As shown in Figure 2, we first evaluate the importance of each parameter based on new task data to generate a fixed gradient mask. Then, we use this mask to protect the majority of model parameters and update only a small subset of parameters most relevant to new task.

#### 3.1. Preliminaries

Continual learning aims to address a key challenge: enabling a model to acquire new tasks sequentially while minimizing performance degradation on previously learned ones. Formally, the model needs to learn a sequence of tasks  $\mathcal{T} = \{T_1, T_2, \dots, T_T\}$ . Each task  $T_t$  corresponds to a dataset  $\mathcal{D}_t = \{(x_i^{(t)}, y_i^{(t)})\}_{i=1}^{n_t}$ , consisting of  $n_t$  input-output pairs. Given a neural network  $f(\cdot; \theta)$  parameterized by  $\theta$ , the objective at each task step  $t$  is to adapt to

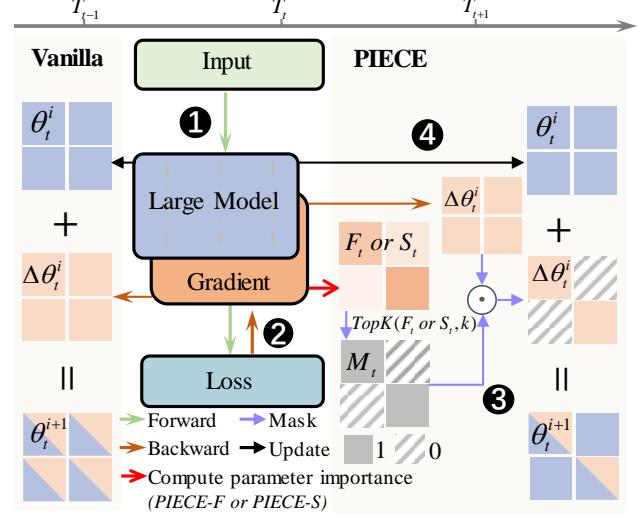


Figure 2. The illustration of PIECE. During parameter updates, PIECE performs standard ① forward and ② backward steps, but before ④ updating, it ③ applies a fixed gradient mask (computed from PIECE-F/S parameter importance) to protect most parameters and update only top-k task-relevant ones.

task  $T_t$  by minimizing:

$$\theta^* = \arg \min_{\theta} \frac{1}{n_t} \sum_{i=1}^{n_t} L(f(x_i^{(t)}; \theta), y_i^{(t)}), \quad (1)$$

where  $L(\cdot, \cdot)$  denotes the loss function, e.g., cross-entropy.

Typically, only the current dataset  $\mathcal{D}_t$  is accessible during the training of task  $T_t$ . However, previous approaches often relax this constraint to varying degrees. For example, replay-based methods retain a small subset of past samples for replay or generate pseudo-samples; regularization-based methods store historical signals such as gradients or parameter importance to constrain updates; architecture-based methods adapt to new tasks by adding parameters or modifying the structure, sometimes with additional task-identification memory. In this work, we strictly follow a **no-history assumption**: while learning task  $T_t$ , the model is prohibited from accessing any past data or intermediate training information, and both the parameter budget and architecture remain fixed.

Moreover, unlike conventional settings that start from task  $T_1$ , we consider a more realistic large-model scenario in which the model may already exhibit strong performance on some tasks. These tasks should be treated as prior knowledge to be preserved. Thus, continual learning can **start from any step**  $t \in \{1, \dots, T\}$  while protecting previously acquired capabilities.

#### 3.2. Fisher Information-based Importance

We employ Fisher Information [16] to assess the importance of model parameters, which is widely used in modern

machine learning for various purposes, such as protecting critical parameters of previous tasks [24], guiding model compression and pruning [49], and selecting target parameters for efficient fine-tuning [50, 58]. From the perspective of parameter importance, a parameter that strongly influences the model’s output will induce a significant change in the predictive distribution when updated. Let  $\theta$  denote the model parameters and  $x$  the input, with the output distribution denoted by  $p_\theta(y|x)$ . When a small perturbation  $\delta \rightarrow 0$  is applied to the parameters (as in the fine-tuning stage), it can be shown [39, 43] that the resulting change in the predictive distribution can be approximated as:

$$\mathbb{E}_x[D_{\text{KL}}(p_\theta(y|x) \| p_{\theta+\delta}(y|x))] \approx \delta^\top F_\theta \delta, \quad (2)$$

where  $F_\theta$  denotes the Fisher information matrix, which is defined as:

$$F_\theta = \mathbb{E}_x \left[ \mathbb{E}_{y \sim p_\theta(y|x)} \left[ \nabla_\theta \log p_\theta(y|x) \nabla_\theta \log p_\theta(y|x)^\top \right] \right]. \quad (3)$$

Since the true data distribution is typically unavailable and computing the full Fisher information matrix is expensive, we follow the approaches of Kirkpatrick et al. [24], Sung et al. [50], and Xu et al. [58] to approximate the importance of each parameter using the diagonal elements of the empirical Fisher information matrix computed over the training set  $D_t$  of a given task. Specifically, for task  $T_t$ , the Fisher Information parameter importance of the  $i$ -th parameter is defined as:

$$F_{t,i} \approx \frac{1}{|D_t|} \sum_{(x,y) \in D_t} \left( \frac{\partial \log p_{\theta_{t-1}}(y|x)}{\partial \theta_{t-1,i}} \right)^2. \quad (4)$$

### 3.3. Second-order Normalization Importance

From a Bayesian perspective, we analyze the rationale of using the Fisher information to measure parameter importance for past tasks, taking the optimization of task  $T_2$  as an example:

$$p(\theta_2 | D_1, D_2) = \frac{p(D_2 | \theta_2)p(\theta_2 | D_1)}{p(D_2)}. \quad (5)$$

Here, the posterior distribution  $p(\theta_2 | D_1)$  from the previous task  $T_1$ , serves as the prior when optimizing the new task, thereby protecting knowledge acquired from past parameters. Maximizing this posterior is equivalent to minimizing the new task loss  $L_2(\theta_2)$  augmented with a regularization term induced by the prior from the old task:

$$\begin{aligned} \theta_2 &= \arg \max_{\theta_2} \log p(\theta_2 | D_1, D_2) \\ &= \arg \min_{\theta_2} L_2(\theta) - \log p(\theta_2 | D_1). \end{aligned} \quad (6)$$

Then perform a second-order Taylor expansion of  $f(\theta_2) = \log p(\theta_2 | D_1)$  around the optimum  $\theta_1$  of the previous task:

$$f(\theta_2) \approx f(\theta_1) + J_{\theta_2, \theta_1}(\theta_2 - \theta_1) + \frac{1}{2}(\theta_2 - \theta_1)^\top H_{\theta_2, \theta_1}(\theta_2 - \theta_1), \quad (7)$$

$$J_{\theta_2, \theta_1} = \left. \frac{\partial f(\theta_2)}{\partial \theta_2} \right|_{\theta_2=\theta_1}, \quad H_{\theta_2, \theta_1} = \left. \frac{\partial^2 f(\theta_2)}{\partial \theta_2^2} \right|_{\theta_2=\theta_1}.$$

Since  $\theta_i$  is a stationary point for the previous task, the Jacobian vanishes, i.e.,  $J_{\theta_2, \theta_1} = 0$ . Moreover, the negative expected Hessian, i.e.,  $H_{\theta_2, \theta_1}$  equals the Fisher information matrix. Thus, using Fisher information to quantify parameter importance is consistent with Bayesian formulation.

Although the Fisher information characterizes parameter importance around the optimum of the previous task, it relies on the assumption that the current parameters remain close to that optimum, where the gradients vanish. However, when learning a new task  $T_t$ , the initialization  $\theta_{t-1}$  is in general not a stationary point under the new data distribution, i.e.,  $\nabla_{\theta_{t-1}} \log p_{\theta_{t-1}}(D_t) \neq 0$ . Neglecting the first-order term thus ignores the primary source of parameter updates for the new task. To address this, we instead expand the new task log-likelihood  $f_t^*(\theta_t) = \log p(\theta_t | D_t)$  at the current parameter  $\theta_{t-1}$ :

$$\begin{aligned} f_t^*(\theta_t) &\approx f_t^*(\theta_{t-1}) + J_{\theta_t, \theta_{t-1}}(\theta_t - \theta_{t-1}) \\ &\quad + \frac{1}{2}(\theta_t - \theta_{t-1})^\top H_{\theta_t, \theta_{t-1}}(\theta_t - \theta_{t-1}), \end{aligned} \quad (8)$$

and yield the approximate optimal update:

$$\theta_t^* \approx \theta_{t-1} - H_{\theta_t, \theta_{t-1}}^{-1} J_{\theta_t, \theta_{t-1}}. \quad (9)$$

This expression indicates that the magnitude of parameter updates is jointly determined by the first-order gradient and the second-order curvature: parameters with large gradients and low curvature tend to change more during learning and thus contribute more to adaptation. Further, assuming that the parameters are near a local optimum, the Laplace approximation gives  $\theta_t^* \sim \mathcal{N}(\theta_t^*, H_{\theta_t^*, \theta_{t-1}}^{-1})$ , while the current parameters  $\theta_{t-1}$  can be viewed as a degenerate distribution  $\delta(\theta_{t-1})$ . We therefore use the standardized mean shift between these two distributions as the importance measure for each parameter. To avoid computing true data expectations and Hessians, we approximate the Jacobian by the empirical gradients and the negative Hessian by the empirical Fisher information computed at  $\theta_{t-1}$  under task  $T_t$ . Specifically, for task  $T_t$  the Second-order Normalized parameter importance of the  $i$ -th parameter is defined as:

$$S_{t,i} = \frac{\left| \frac{1}{|D_t|} \sum_{(x,y) \in D_t} \frac{\partial \log p_{\theta_{t-1}}(y|x)}{\partial \theta_{t-1,i}} \right|}{\sqrt{\frac{1}{|D_t|} \sum_{(x,y) \in D_t} \left( \frac{\partial \log p_{\theta_{t-1}}(y|x)}{\partial \theta_{t-1,i}} \right)^2} + \xi}, \quad (10)$$

where  $\xi > 0$  is a constant for numerical stability.

### 3.4. PIECE

We introduce the core workflow of PIECE using the most standard single-step fine-tuning paradigm as an example.

Let  $\theta_t^i$  denote the model parameters at the  $i$ -th iteration on task  $T_t$  (where  $\theta_t^0$  denotes the initialization for task  $T_t$ ). In standard fine-tuning, gradients of the loss  $L(f(x^{(t)}; \theta_t^i), y^{(t)})$  are computed and applied to update all parameters via gradient descent:

$$\theta_t^{i+1} = \theta_t^i - \eta \frac{\partial L(f(x^{(t)}; \theta_t^i), y^{(t)})}{\partial \theta_t^i}. \quad (11)$$

In contrast, PIECE first computes the importance of each parameter from the current task data (as defined in Eq. 4 or Eq. 10), and only updates the top- $k$  most important parameters, forming a task-specific sparse update subnetwork  $\mathcal{C}_t^k$ . In practice, we set  $k = \lfloor 0.001 \cdot |\theta| \rfloor$ , i.e., only 0.1% of parameters are trainable for each task. To achieve this, we define a binary mask with the same shape as  $\theta$ :

$$M_{t,i} = \begin{cases} 1, & \theta_{t,i} \in \mathcal{C}_t^k \\ 0, & \theta_{t,i} \notin \mathcal{C}_t^k \end{cases}. \quad (12)$$

It is then applied to filter gradient updates, freezing low-importance parameters and updating the important ones:

$$\theta_t^{i+1} = \theta_t^i - \eta \left( \frac{\partial L(f(x^{(t)}; \theta_t^i), y^{(t)})}{\partial \theta_t^i} \odot M_t \right). \quad (13)$$

Algorithm 1 provides the complete pseudo-code of PIECE in a sequential multi-task learning setting.

## 4. Experiments

We validate our method on several benchmark datasets against a variety of continual learning baselines.

### 4.1. Experiment Setup

**Datasets** For language tasks, we use the TRACE [57] benchmark with 1,000 training samples per task. TRACE contains eight sub-datasets spanning science (Sin.), finance (Fom.), social news (Mee.), multiple languages (English, German [20M.], Chinese [Cst.]), and complex tasks such as code questions (Py150) and mathematical reasoning (Ngc, Ngd). We use the HumanEval [9] programming benchmark to assess the model’s capability retention. For multimodal tasks, we use VQA v2 [19]. Following VQACL [63], tasks are split by question type with 1,000 training samples per task, keeping the original test set. We focus on three types—action (act.), commonsense (com.), and count (cou.)—based on performance gains and capability preservation. Flickr30K [61] is used to assess overall capability retention. Details in supplementary materials A.

**Baselines** We compared three categories of baselines: **Regularization** methods like EWC [24], GEM [34], and LwF [27] constrain updates via parameter importance, gradient projection, or distillation to preserve prior knowledge. **Replay** methods reuse past data to reduce forgetting,

---

**Algorithm 1** PIECE: Parameter Importance Estimation-based Continual Enhancement

---

```

Require: Sequential tasks  $\{T_1, T_2, \dots, T_T\}$ ; model  $f(\cdot; \theta)$ ;
          top- $k$  ratio  $k$ ; learning rate  $\eta$ ;
1: for  $t = 1$  to  $T$  do
2:   Initialize  $\theta_t^0 = \theta_{t-1}$ 
   //  $\theta_0$  is foundation model parameters.
3:   Sample dataset  $D_t$ 
4:   for each parameter  $\theta_{t-1,i}$  do
5:     Compute importance score  $F_{t,i}$  or  $S_{t,i}$ 
     via Eq. 4 or Eq. 10
6:   end for
   // Compute parameter importance for task  $T_t$ 
7:    $\mathcal{C}_t^k = \text{TopK}(F_t \text{ or } S_t, k)$ 
8:   Construct binary mask  $M_t$  where  $M_{t,i} = 1$  if  $\theta_{t,i} \in \mathcal{C}_t^k$ , else 0
   // Build sparse update subnetwork
   //  $I_t$ : number of optimization steps for task  $T_t$ 
9:   for  $i = 0$  to  $I_t$  do
10:    Compute gradient  $g = \frac{\partial L(f(x^{(t)}; \theta_t^i), y^{(t)})}{\partial \theta_t^i}$ 
11:     $\tilde{g} = g \odot M_t$ 
12:     $\theta_t^{i+1} = \theta_t^i - \eta \tilde{g}$ 
13:   end for
   // Gradient descent on selected parameters
14: end for
15: return  $\theta_T$ 

```

---

either after each new task (Replay) [44] or continuously during training (Replay-online) [45]. **PET** baselines include SeqLoRA [21], O-LoRA [56], LayerNorm [65], and MIGU [14], which adapt different subsets or structures of parameters. Full fine-tuning (SeqFT) is also included. Details in supplementary materials B.

**Metrics** Let  $R_{t,i}$  denote the model’s performance on task  $i$  after learning task  $t$ . Accuracy is used for classification or single-token QA tasks, and the average of Rouge-L and BLEU scores for others. The evaluation metrics are Overall Performance (OP) [7]  $OP_t = \frac{1}{t} \sum_{i=1}^t R_{t,i}$ , measuring average performance across all learned tasks; Backward Transfer (BWT) [34]  $BWT_t = \frac{1}{t} \sum_{i=1}^{t-1} (R_{t,i} - R_{i,i})$ , indicating how later tasks affect earlier ones. These are computed on sequential tasks to compare methods and isolate changes in the model’s original abilities. For capability preservation evaluation, code generation (HumanEval) is measured by Pass@1, i.e., the fraction of correct single-attempt outputs, and Image captioning (Flickr30K) is measured by the mean of Rouge-L and BLEU scores. See supplementary material C for details.

**Models and Implementation Details** We use Gemma2-2B [51], Llama3-8B [1], and Qwen3-14B [52] for language tasks, and Qwen3-VL-4B [52] and LLaVA-1.5-7B [32, 33] for multimodal tasks. Language tasks follow the TRACE

Methods	HumanEval(↑)	Various task in TRACE (↑)								<i>OP</i> (↑)	<i>BWT</i> (↑)
		Cst.	Fom.	Mee.	Sin.	Ngc.	Ngd.	20M.	Py150		
<i>Gemma2-2B</i>	30.49										
SeqFT	0.0	23.70	0.0	7.28	1.65	7.40	38.77	2.91	39.35	15.13	-19.08
EWC	0.0	21.25	18.95	5.36	3.50	2.47	39.69	3.21	42.65	17.13	-13.99
GEM	0.0	32.95	13.51	4.50	16.20	6.17	35.69	2.86	41.31	19.15	-16.27
LwF	0.0	10.15	11.41	11.64	10.01	17.41	39.69	10.24	14.33	16.86	-10.10
replay	0.0	33.40	37.09	17.19	43.15	3.70	32.62	7.96	34.01	26.14	-1.43
replay-online	0.0	33.10	40.52	18.52	44.45	13.58	34.77	8.53	39.35	29.10	-3.82
SeqLoRA	10.37	45.75	35.28	11.31	67.25	<b>38.27</b>	50.77	12.18	<b>57.36</b>	39.77	-10.80
O-LoRA	13.41	48.85	<b>57.26</b>	16.24	72.00	33.33	<b>51.38</b>	11.57	54.55	43.15	-3.45
LayerNorm	21.95	52.75	35.28	8.65	75.75	12.35	9.54	9.78	13.78	27.23	-1.67
MIGU	1.22	35.00	4.64	9.46	16.00	12.35	42.77	8.33	53.12	22.71	-10.36
PIECE-F (ours)	32.32	53.10	53.43	18.17	<b>76.25</b>	34.57	50.77	<b>12.90</b>	55.48	<b>44.33</b>	-3.21
PIECE-S (ours)	<b>33.54</b>	<b>54.35</b>	52.22	<b>19.37</b>	75.85	22.22	45.23	12.75	54.24	42.03	<b>-1.18</b>
<i>Llama3-8B</i>	34.15										
SeqFT	0.0	47.10	58.67	23.77	41.75	28.40	47.70	13.32	53.05	39.22	-10.69
EWC	0.0	33.95	<b>58.67</b>	22.10	21.20	27.16	39.08	11.33	50.13	32.95	-10.05
GEM	0.0	41.05	<b>60.48</b>	24.33	47.75	27.16	48.92	10.80	54.43	39.37	-10.24
LwF	0.0	39.80	44.07	12.14	40.20	23.45	49.23	10.67	44.92	33.06	-9.94
replay	0.0	42.45	<b>60.48</b>	25.88	53.55	25.93	40.62	11.20	48.84	38.62	-2.94
replay-online	0.0	<b>52.25</b>	<b>65.32</b>	<b>26.46</b>	<b>76.95</b>	38.27	42.46	10.78	51.25	45.47	-1.99
SeqLoRA	28.05	47.15	56.05	18.62	72.75	54.32	<b>53.23</b>	12.51	53.67	46.04	-4.39
O-LoRA	26.83	45.20	39.52	17.57	61.30	43.20	48.00	11.73	51.30	39.73	-7.19
LayerNorm	27.44	34.90	27.02	9.44	66.80	37.04	35.08	9.34	23.50	30.39	-1.86
MIGU	0.0	38.60	25.40	20.93	59.35	41.98	49.23	11.42	54.29	37.65	-10.02
PIECE-F (ours)	29.88	47.95	59.88	23.42	76.40	<b>61.73</b>	<b>57.85</b>	13.50	<b>59.65</b>	<b>50.05</b>	-2.66
PIECE-S (ours)	<b>32.32</b>	49.05	57.66	23.47	<b>77.10</b>	59.26	52.62	<b>14.00</b>	59.24	49.05	<b>-1.78</b>
<i>Qwen3-14B</i> *	59.76										
SeqFT	47.56	13.30	4.84	8.86	0.0	25.91	24.00	<b>12.64</b>	26.88	14.56	-16.24
EWC	46.34	38.95	39.72	12.79	86.65	40.74	35.69	11.86	26.85	36.66	-8.27
GEM	48.17	37.85	43.15	13.64	86.75	50.62	48.00	11.33	20.93	39.03	-7.96
LwF	<b>54.27</b>	18.60	0.20	7.45	44.60	41.98	25.85	1.78	22.18	20.33	-2.04
replay	45.12	42.60	50.81	<b>30.60</b>	65.65	33.33	8.92	12.04	<b>43.52</b>	35.93	-7.64
replay-online	48.17	7.25	0.0	<b>25.91</b>	85.25	49.38	<b>49.54</b>	12.58	41.71	33.95	-8.89
SeqLoRA	49.39	26.90	55.04	12.13	64.80	<b>58.02</b>	44.62	11.70	53.21	40.80	-6.31
O-LoRA	44.51	47.05	53.83	10.40	89.25	55.56	<b>51.38</b>	10.63	27.73	43.23	-2.20
LayerNorm	50.00	<b>63.50</b>	40.93	9.50	44.25	0.0	0.62	10.48	4.10	21.67	-0.76
MIGU	49.39	42.60	39.72	15.41	11.45	41.98	35.69	<b>13.01</b>	19.82	31.38	-7.27
PIECE-F (ours)	50.61	<b>60.65</b>	<b>61.49</b>	9.41	<b>90.55</b>	48.15	44.62	10.63	23.15	<b>49.80</b>	-0.33
PIECE-S (ours)	<b>59.14</b>	59.45	<b>62.50</b>	13.39	<b>90.50</b>	<b>62.96</b>	46.77	11.96	<b>47.27</b>	49.35	<b>-0.16</b>

Table 1. Performance of methods on language tasks. OP and BWT are computed on sequential tasks; inherent capability variation is measured via HumanEval. \*Qwen3-14B trained one epoch per task, showing less forgetting.

order, with Py150 placed last to limit its impact on intrinsic capabilities; multimodal tasks are sequenced by their effect on original capabilities: action, commonsense, count. All models use the Adam optimizer (batch size 64, seed

42), with learning rates: Gemma2-2B, 5e-4 for LoRA/O-LoRA, and 5e-5 for the others; for the other models, 1e-4 for LoRA/O-LoRA, and 1e-5 for the rest. No warmup is applied. Qwen3-14B is trained 1 epoch per task, others 5

Methods	Flickr30k( $\uparrow$ )	Various task in VQA			$OP(\uparrow)$	$BWT(\uparrow)$
		Act.	Com.	Cou.		
<i>Qwen3-VL-4B</i>	<b>26.19</b>					
SeqFT	7.42	47.29	48.66	44.03	46.66	-13.86
EWC	4.34	40.06	40.42	38.40	39.62	-18.14
GEM	8.28	47.64	42.54	41.34	43.84	-12.89
LwF	<b>26.67</b>	<b>66.97</b>	58.19	40.89	55.35	-5.74
replay	20.79	61.20	60.88	<b>52.27</b>	<b>58.11</b>	-2.72
replay-online	8.86	<u>65.65</u>	<b>69.68</b>	<u>50.67</u>	62.00	-0.83
SeqLoRA	<b>32.18</b>	38.66	45.56	23.38	35.87	-1.62
O-LoRA	5.23	40.06	48.90	37.01	41.99	-1.25
LayerNorm	<b>23.77</b>	60.29	59.09	43.06	<b>54.15</b>	-1.37
MIGU	27.44	47.64	57.46	37.29	47.46	-4.87
PIECE-F (ours)	35.59	62.66	58.76	48.19	<b>56.54</b>	-1.53
PIECE-S (ours)	<b>38.74</b>	62.15	<u>61.53</u>	47.50	57.06	<b>-0.50</b>
<i>LLaVA-1.5-7B</i>	<b>30.82</b>					
SeqFT	10.69	41.72	39.36	43.03	41.37	-26.84
EWC	21.92	46.38	52.40	24.21	41.00	-10.04
GEM	24.17	49.65	53.63	30.16	44.48	-7.18
LwF	<b>36.02</b>	67.66	58.11	2.07	42.62	-0.51
replay	13.61	65.79	70.90	45.62	<b>60.77</b>	-0.78
replay-online	10.46	67.25	69.11	44.93	60.43	-1.66
SeqLoRA	<b>31.04</b>	<b>75.31</b>	75.63	<u>53.10</u>	68.01	-0.51
O-LoRA	6.80	70.38	72.94	52.30	<b>65.21</b>	-0.26
LayerNorm	<b>30.71</b>	69.68	60.39	2.91	44.33	-0.20
MIGU	21.81	67.66	52.40	42.89	<b>54.32</b>	-8.62
PIECE-F (ours)	37.51	<u>74.48</u>	<u>77.18</u>	53.06	68.24	-0.29
PIECE-S (ours)	<b>41.94</b>	<b>75.31</b>	<b>77.34</b>	<b>53.27</b>	<b>68.64</b>	<b>-0.16</b>

Table 2. Results of different methods on multimodal tasks.

epochs. Experiments run on 4x80G A100, 4x80G A800, and 8x40G A100 GPUs. See supplementary material B for baseline-specific deployment details.

## 4.2. Main Results

The main results are shown in Table 1 and Table 2. Table 1 presents language tasks performance, and Table 2 reports multimodal results. Based on these results, we can draw the following conclusions:

**PIECE demonstrates outstanding performance across both language and multimodal tasks.** Across models of different scales, especially for large models, both PIECE-F and PIECE-S achieve the highest OP, outperforming all baselines. Meanwhile, the BWT results indicate that PIECE substantially alleviates forgetting during sequential task learning, highlighting its strong potential to balance new task acquisition and previous knowledge retention.

**PIECE effectively preserves the original capabilities of foundation models.** On the HumanEval benchmark for language models and the Flickr30K benchmark for multimodal models, PIECE significantly outperforms traditional approaches, verifying its robustness in maintaining general capabilities. Even for Qwen3-14B with one-epoch-per-task training and naturally mild forgetting, PIECE still maintains a clear advantage, demonstrating its consistent ability-preserving effect under different training intensities.

**PIECE-F excels in task transfer, while PIECE-S**

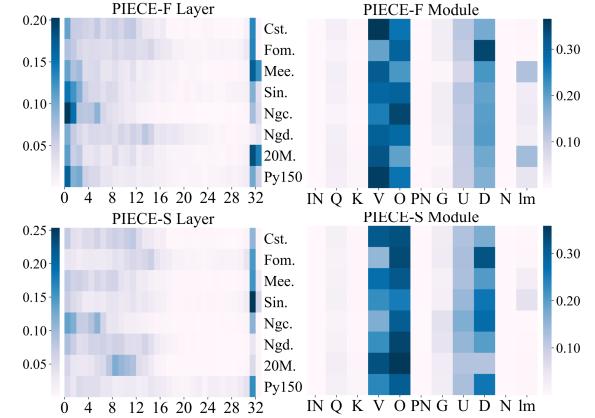


Figure 3. Distribution of critical parameters identified by **PIECE-F** and **PIECE-S** across different tasks.

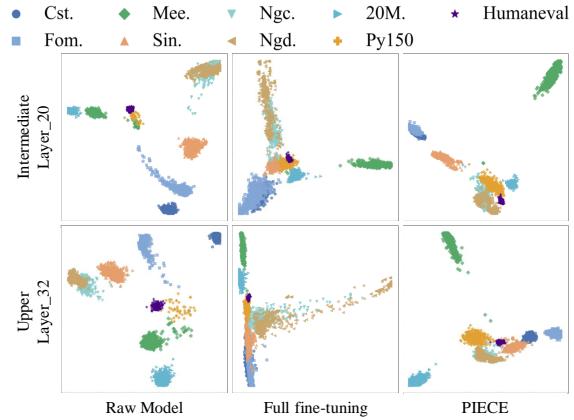


Figure 4. Visualization of intermediate and upper layer task representations comparing the base model, full fine-tuning, and **PIECE**.

**achieves stronger forgetting mitigation.** As reflected by the OP metric, PIECE-F slightly outperforms PIECE-S across several tasks, showing greater adaptability and transferability to new domains. Conversely, PIECE-S consistently achieves the best BWT results with the least forgetting. This stability not only mitigates performance degradation in sequential learning but also further contributes to preserving the foundation model’s original abilities.

## 4.3. Analysis

To further understand the behavior of **PIECE** beyond its overall performance, we conduct a series of analytical experiments focusing on three key questions:

1. *Where PIECE learns and why it achieves its objective?*
2. *Whether PIECE forgets more as task count increases?*
3. *How the choice of sparsity ratio affects performance?*

All analyses are performed on the **Llama3-8B** model.

*Where PIECE learns and why it achieves its objective?*

To explore where **PIECE** learns and why it is stable, we visualize the important parameters selected by **PIECE-F**

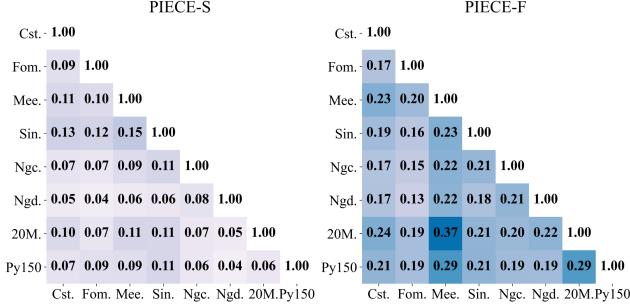


Figure 5. Parameter overlap across tasks.

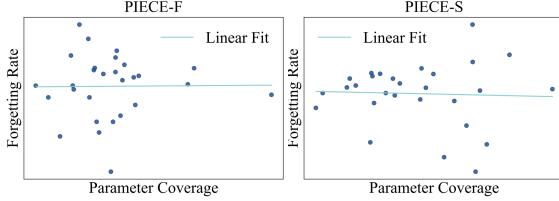


Figure 6. Correlation of parameter overlap and forgetting rate.

and **PIECE-S** across tasks. Figure 3 shows both methods mainly target attention submodules V/O and feed-forward layers (U/D). This aligns with Yao et al. [60], which suggests that the V/O modules play a more crucial role than Q/K in downstream adaptation, and Dai et al. [12], which identifies U and D as key locations for knowledge storage. At the layer level, both methods primarily attend to the higher layers, consistent with Zhao et al. [68] linking them to downstream task relevance. The distinction is that **PIECE-F** also emphasizes lower layers, potentially aiding better adaptation to downstream tasks, although adjusting these shared modules may introduce some forgetting. Notably, both mechanisms avoid the intermediate layers, suggesting that these layers encode core pretraining knowledge and are crucial for maintaining the structural stability of the model. To further examine the role of intermediate layers in the model’s multitask capabilities, we visualized task representations at intermediate and upper layers using dimensionality reduction. As shown in Figure 4, compared to base model, full fine-tuning disrupts the representations in intermediate layers, consequently impairing the final task separability, whereas **PIECE** preserves clear task distinctions.

#### Whether **PIECE** forgets more as task count increases?

The core of **PIECE** lies in fine-tuning only a small subset of critical parameters, thereby mitigating catastrophic forgetting. This raises a natural question: as the number of tasks increases, do important parameters overlap significantly across tasks, potentially increasing the risk of forgetting? To investigate the scalability of **PIECE**, we first analyze the parameter overlap across tasks. As shown in Figure 5, the overlap is relatively low for both **PIECE-F** and **PIECE-S**, with **PIECE-S**, which emphasizes task-specific parameters, exhibiting the lowest overlap. Moreover, we

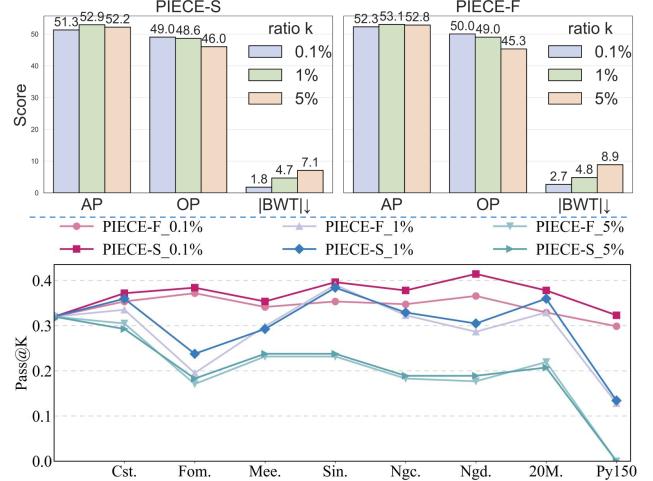


Figure 7. Effect of the number of updated parameters on performance and forgetting. Average Performance (AP) is the mean performance across sequential tasks after each task, reflecting transfer ability:  $AP_t = \frac{1}{t} \sum_{i=1}^t R_{i,i}$

compute the correlation between parameter overlap and task forgetting rate (Figure 6). The results show negligible correlation for both methods (Pearson r: -0.04 for **PIECE-S**, 0.01 for **PIECE-F**), indicating that even as the number of tasks grows, **PIECE** can effectively preserve critical parameters and maintain performance on previous tasks.

#### How the choice of sparsity ratio affects performance?

To investigate the effect of the number of updated parameters on performance, we varied the selection ratio  $k$ , with the results shown in Figure 7. The experiments indicate that increasing  $k$  initially improves the model’s transfer performance on downstream tasks, but this improvement is not sustained; at  $k = 5\%$ , performance is actually lower than at  $k = 1\%$ . Meanwhile, larger  $k$  values substantially exacerbate forgetting, resulting in a decrease in OP and causing more severe degradation of the original capabilities.

## 5. Conclusion

We present **PIECE**, a Parameter Importance Estimation-based Continual Enhancement method for foundation models. **PIECE** mitigates catastrophic forgetting by selectively updating only the most critical 0.1% of parameters, guided by theoretically grounded importance estimation. Without relying on replay data, task labels, or architectural modifications, **PIECE** achieves efficient domain adaptation while preserving general reasoning and programming abilities. Extensive experiments on diverse LLMs and MLLMs show that **PIECE** surpasses existing approaches in both retention and transfer. By coupling parameter-efficient tuning with importance estimation, **PIECE** offers a scalable, model-agnostic pathway toward sustainable, domain-adaptive foundation models. In the future, we will fur-

ther distinguish fine-tuning key parameters from stability-critical ones to better balance downstream transfer and forgetting mitigation.

## 6. Acknowledgments

This work was funded by the National Natural Science Foundation of China (NSFC) under Grants No. 62406013, the Beijing Advanced Innovation Center Funds for Future Blockchain and Privacy Computing(GJJ-24-034), and the Fundamental Research Funds for the Central Universities.

## References

- [1] AI@Meta. Llama 3 model card. 2024. [5](#)
- [2] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3366–3375, 2017. [3](#)
- [3] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018. [2](#)
- [4] Claas Beger and Saikat Dutta. Coconut: Structural code understanding does not fall out of a tree. In *2025 IEEE/ACM International Workshop on Large Language Models for Code (LLM4Code)*, pages 128–136. IEEE, 2025. [1](#)
- [5] Prashant Shivaram Bhat, Shakib Yazdani, Elahe Arani, and Bahram Zonooz. Parameter efficient continual learning with dynamic low-rank adaptation. *arXiv preprint arXiv:2505.11998*, 2025. [3](#)
- [6] Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. Lora learns less and forgets less. *Transactions on Machine Learning Research*, 2024. [3](#)
- [7] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–547, 2018. [5](#)
- [8] Jie Chen, Zhipeng Chen, Jiapeng Wang, Kun Zhou, Yutao Zhu, Jinhao Jiang, Yingqian Min, Wayne Xin Zhao, Zhicheng Dou, Jiaxin Mao, et al. Towards effective and efficient continual pre-training of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5779–5795, 2025. [2, 1](#)
- [9] Mark Chen. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. [5](#)
- [10] Qian Chen, Lei Zhu, Hangzhou He, Xinliang Zhang, Shuang Zeng, Qiushi Ren, and Yanye Lu. Low-rank mixture-of-experts for continual medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 382–392. Springer, 2024. [3](#)
- [11] Yanyuan Chen, Dexuan Xu, Yu Huang, Songkun Zhan, Hanpin Wang, Dongxue Chen, Xueping Wang, Meikang Qiu, and Hang Li. Mimo: A medical vision language model with visual referring multimodal input and pixel grounding multimodal output. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24732–24741, 2025. [1](#)
- [12] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, 2022. [8](#)
- [13] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature machine intelligence*, 5(3):220–235, 2023. [3](#)
- [14] Wenyu Du, Shuang Cheng, Tongxu Luo, Zihan Qiu, Zeyu Huang, Ka Chun Cheung, Reynold Cheng, and Jie Fu. Unlocking continual learning abilities in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6503–6522, 2024. [2, 3, 5, 1](#)
- [15] Yujie Feng, Xujia Wang, Zexin Lu, Shenghong Fu, Guangyuan Shi, Yongxin Xu, Yasha Wang, Philip S Yu, Xu Chu, and Xiao-Ming Wu. Recurrent knowledge identification and fusion for language model continual learning. *arXiv preprint arXiv:2502.17510*, 2025. [1](#)
- [16] Ronald Aylmer Fisher. Theory of statistical estimation. In *Mathematical proceedings of the Cambridge philosophical society*, pages 700–725. Cambridge University Press, 1925. [3](#)
- [17] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999. [1](#)
- [18] Annette Rios Gonzales, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. A new dataset and efficient baselines for document-level text simplification in german. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 152–161, 2021. [1](#)
- [19] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. [5, 2](#)
- [20] Jinghan He, Haiyun Guo, Kuan Zhu, Zihan Zhao, Ming Tang, and Jinqiao Wang. Seekr: Selective attention-guided knowledge retention for continual learning of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3254–3266, 2024. [1](#)
- [21] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*. [2, 3, 5](#)
- [22] Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. Meetingbank:

- A benchmark dataset for meeting summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16409–16423, 2023. 1
- [23] Nikitas Karanikolas, Eirini Manga, Nikoletta Samaridi, Eleni Tousidou, and Michael Vassilakopoulos. Large language models versus natural language understanding and generation. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, pages 278–290, 2023. 1
- [24] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2, 4, 5
- [25] Minh Le, An Nguyen, Huy Nguyen, Trang Nguyen, Trang Pham, Linh Van Ngo, and Nhat Ho. Mixture of experts meets prompt-based continual learning. *Advances in Neural Information Processing Systems*, 37:119025–119062, 2024. 3
- [26] Jia Li, Ge Li, Yongmin Li, and Zhi Jin. Structured chain-of-thought prompting for code generation. *ACM Transactions on Software Engineering and Methodology*, 34(2):1–23, 2025. 1
- [27] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 2, 5
- [28] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Benchmark evaluations and challenges. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1587–1606, 2025. 1
- [29] Huanxuan Liao, Shizhu He, Yupu Hao, Jun Zhao, and Kang Liu. Data: Decomposed attention-based task adaptation for rehearsal-free continual learning. *arXiv preprint arXiv:2502.11482*, 2025. 2
- [30] Chengyuan Liu, Yangyang Kang, Shihang Wang, Lizhi Qing, Fubang Zhao, Chao Wu, Changlong Sun, Kun Kuang, and Fei Wu. More than catastrophic forgetting: Integrating general capabilities for domain-specific llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7531–7548, 2024. 2
- [31] Hong Liu, Mingsheng Long, Jianmin Wang, and Yu Wang. Learning to adapt to evolving domains. *Advances in neural information processing systems*, 33:22338–22348, 2020. 1
- [32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 5
- [33] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 5
- [34] David Lopez-Paz and Marc’Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017. 2, 5
- [35] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 1
- [36] Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, et al. Codexglue: A machine learning benchmark dataset for code understanding and generation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. 1
- [37] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 2025. 2
- [38] Aru Maekawa, Hidetaka Kamigaito, Kotaro Funakoshi, and Manabu Okumura. Generative replay inspired by hippocampal memory indexing for continual language learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 930–942, 2023. 2
- [39] James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. 4
- [40] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, pages 109–165. Elsevier, 1989. 2
- [41] Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. Numglue: A suite of fundamental yet challenging mathematical reasoning tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3505–3523. Association for Computational Linguistics, 2022. 1
- [42] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431, 2024. 1
- [43] Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. *arXiv preprint arXiv:1301.3584*, 2013. 4
- [44] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 2, 5
- [45] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019. 2, 5
- [46] Agam Shah, Suvan Paturi, and Sudheer Chava. Trillion dollar words: A new financial dataset, task & market analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6664–6679, 2023. 1

- [47] Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 2024. 2
- [48] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017. 2
- [49] Sidak Pal Singh and Dan Alistarh. Woodfisher: Efficient second-order approximation for neural network compression. *Advances in Neural Information Processing Systems*, 33:18098–18109, 2020. 4
- [50] Yi-Lin Sung, Varun Nair, and Colin A Raffel. Training neural networks with fixed sparse masks. *Advances in Neural Information Processing Systems*, 34:24193–24205, 2021. 2, 3, 4
- [51] Gemma Team. Gemma. 2024. 5
- [52] Qwen Team. Qwen3 technical report, 2025. 5
- [53] Kai Tong, Kang Pan, Xiao Zhang, Erli Meng, Run He, Yawen Cui, Nuoyan Guo, and Huiping Zhuang. Any-ssr: How recursive least squares works in continual learning of large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3047–3057, 2025. 2, 3, 1
- [54] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2
- [55] Vinay Kumar Verma, Kevin J Liang, Nikhil Mehta, Piyush Rai, and Lawrence Carin. Efficient feature transformations for discriminative and generative continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13865–13875, 2021. 2
- [56] Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuan-Jing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10658–10671, 2023. 2, 3, 5
- [57] Xiao Wang, Yuansen Zhang, Tianze Chen, Songyang Gao, Senjie Jin, Xianjun Yang, Zhiheng Xi, Rui Zheng, Yicheng Zou, Tao Gui, et al. Trace: A comprehensive benchmark for continual learning in large language models. *arXiv preprint arXiv:2310.06762*, 2023. 5
- [58] Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. Raise a child in large language model: Towards effective and generalizable fine-tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9514–9528, 2021. 2, 3, 4
- [59] Yibo Yan, Shen Wang, Jiahao Huo, Jingheng Ye, Zhendong Chu, Xuming Hu, Philip S. Yu, Carla Gomes, Bart Selman, and Qingsong Wen. Position: Multimodal large language models can significantly advance scientific reasoning. *arXiv preprint arXiv:2502.02871*, 2025. 1
- [60] Xinhao Yao, Hongjin Qian, Xiaolin Hu, Gengze Xu, Wei Liu, Jian Luan, Bin Wang, and Yong Liu. Theoretical insights into fine-tuning attention mechanism: Generalization and optimization. *arXiv preprint arXiv:2410.02247*, 2024. 8
- [61] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 5, 2
- [62] Dan Zhang, Tao Feng, Lilong Xue, Yuandong Wang, Yuxiao Dong, and Jie Tang. Parameter-efficient fine-tuning for foundation models. *arXiv preprint arXiv:2501.13787*, 2025. 3
- [63] Xi Zhang, Feifei Zhang, and Changsheng Xu. Vqacl: A novel visual question answering continual learning setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19102–19112, 2023. 5, 2
- [64] Xinyu Zhang, Yuanquan Hu, Fangchao Liu, and Zhicheng Dou. P3: Prompts promote prompting. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11948–11965, 2025. 1
- [65] Bingchen Zhao, Haoqin Tu, Chen Wei, Jieru Mei, and Ci-hang Xie. Tuning layernorm in attention: Towards efficient multi-modal llm finetuning. In *ICLR*, 2024. 2, 5, 3
- [66] Chenye Zhao, Yingjie Li, and Cornelia Caragea. C-stance: A large dataset for chinese zero-shot stance detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13369–13385, 2023. 1
- [67] Weixiang Zhao, Shilong Wang, Yulin Hu, Yanyan Zhao, Bing Qin, Xuanyu Zhang, Qing Yang, Dongliang Xu, and Wanxiang Che. Sapt: A shared attention framework for parameter-efficient continual learning of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11641–11661, 2024. 2, 3
- [68] Zheng Zhao, Yftah Ziser, and Shay B Cohen. Layer by layer: Uncovering where multi-task learning happens in instruction-tuned large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15195–15214, 2024. 8
- [69] Huiping Zhuang, Zhenyu Weng, Hongxin Wei, Renchunzi Xie, Kar-Ann Toh, and Zhiping Lin. Acil: Analytic class-incremental learning with absolute memorization and privacy protection. *Advances in Neural Information Processing Systems*, 35:11602–11614, 2022. 1
- [70] Tao Zhuo, Zhiyong Cheng, Zan Gao, Hehe Fan, and Mohan Kankanhalli. Continual learning with strong experience replay. *arXiv preprint arXiv:2305.13622*, 2023. 2

# Parameter Importance-Driven Continual Learning for Foundation Models

## Supplementary Material

### A Datasets

#### A.1 Language Task Details

The language tasks in this paper are derived from the TRACE dataset, which was constructed following three principles:

1. the datasets should be sufficiently novel so that large language models (LLMs) exhibit weak initial performance;
2. the tasks should be challenging enough to comprehensively test reasoning and generalization capabilities;
3. the tasks should cover a wide range of domains and types.

Due to its well-designed and diverse task settings, TRACE has been adopted in several recent continual learning studies [14, 15, 20, 53]. The language tasks in TRACE can be categorized into four groups: **Domain-Specific Tasks**, **Multi-lingual Tasks**, **Code QA Tasks**, and **Mathematical Reasoning Tasks**. The statistics for each category are shown in Table 3, and configurations are introduced below.

#### Domain-Specific Tasks

Domain-specific tasks evaluate a model’s knowledge acquisition and application within specialized domains. When relevant domain information is underrepresented in training corpora, model performance can degrade significantly. TRACE includes the following three datasets under this category:

- **ScienceQA (Sin.)** [35]: Collected from elementary and high school science curricula, covering topics across natural science, social science, and language science. This dataset requires multi-hop reasoning and scientific knowledge understanding. Only text-only samples are retained in this work.
- **FOMC (Fom.)** [46]: A financial-domain classification task aimed at identifying whether Federal Reserve policy statements are “hawkish” or “dovish.” The original corpus consists of three parts—meeting minutes, press conference transcripts, and speeches. A combined version of all three is used in this work.
- **MeetingBank (Mee.)** [22]: A newly introduced dataset for city council meeting summarization, requiring global comprehension and compression of lengthy meeting transcripts. As a relatively unexplored domain, it effectively evaluates models’ long-context understanding and abstraction abilities.

#### Multi-lingual Tasks

Cross-lingual capability serves as an important measure of a large language model’s generalization ability. However, due to vocabulary coverage and corpus imbalance, most models perform worse on non-English languages. TRACE includes

two datasets for evaluating such multi-lingual abilities:

- **C-STANCE (Cst.)** [66]: A Chinese stance detection dataset collected from Sina Weibo. It includes two subtasks: target-based and domain-based stance detection. We adopt the target-based subtask, where targets in the test set do not appear in the training set, to evaluate zero-shot and cross-domain generalization.
- **20Minuten (20M.)** [18]: A German text simplification dataset derived from the Swiss news magazine 20 Minuten. Each instance consists of an original article and its simplified summary, used to evaluate model performance in German text generation.

#### Code QA Tasks

Code-related tasks are used to assess a model’s understanding and generation capabilities under structured, long-context conditions. TRACE employs the Py150 [36] corpus to construct the code QA task. The corpus contains approximately 150,000 Python programs collected from GitHub repositories. The task requires predicting the subsequent content given code context, where label formats differ from those in standard code generation tasks, posing additional challenges to capability retention. Considering that **HumanEval** is used in this paper to measure the retention of code generation ability, this task plays a crucial role in assessing model performance. Thus, we place Py150 last in TRACE testing order to limit its impact on intrinsic capabilities.

#### Mathematical Reasoning Tasks

Mathematical reasoning tasks are designed to test models’ arithmetic and logical reasoning capabilities. TRACE selects the first two subtasks (Ngc., Ngd.) from the **NumGLUE** benchmark [41]. Both subtasks require numerical computation and logical inference in natural language settings. These tasks impose higher demands on symbolic reasoning and are particularly useful for evaluating capability retention under abstract reasoning scenarios.

Given that programming ability is widely recognized as a representative prior knowledge that is particularly susceptible to forgetting during post-training [8], we adopt it as a measure of the model’s retention of original capabilities and quantify it using the **HumanEval** benchmark. HumanEval is a standardized evaluation dataset for Python function generation tasks, designed to assess the code generation capabilities of large language models [4, 26, 64]. The dataset contains 164 programming problems, each consisting of a function description (function docstring) and a function signature. The model is required to generate a complete, executable function implementation based on the natural language description. Unlike typical natural language genera-

Dataset	Language Task								Multi-modal Task				
	Cst.	Fom.	Mee.	Sin.	Ngc.	Ngd.	20M.	Py150.	HumanEval	Act.	Com.	Cou.	Flickr30k
Train	1000								-	1000			-
Test	2000	496	692	2000	81	325	200	2000	164	1448	1227	2905	1000

Table 3. Dataset Statistics for Language and Multi-modal Tasks.

tion tasks, HumanEval evaluation relies on unit tests rather than manual annotation. The dataset is not used for training, and is solely employed to assess the model’s retention of original capabilities after multi-task continual learning. Since programming tasks demand a high degree of logical consistency, semantic accuracy, and structured reasoning, and these abilities are particularly vulnerable to degradation during multi-task training, performance changes on HumanEval provide a sensitive and reliable indicator of knowledge retention and forgetting in continual learning scenarios.

## A.2 Multi-modal Task Details

The multi-modal tasks in this paper are derived from the VQA v2 dataset [19]. To facilitate continual learning experiments, we follow the task grouping scheme of VQACL [63], splitting the tasks by question type and randomly sampling 1,000 training examples per task while retaining the original test sets. Based on performance improvements and capability retention, we focus on the following three question types in this study:

- **Action (Act.)**: Action recognition questions, which require the model to understand the actions of people or objects in an image;
- **Commonsense (Com.)**: Commonsense reasoning questions, which require the model to make judgments based on both visual information and general knowledge;
- **Count (Cou.)**: Counting questions, which require the model to accurately identify the number of target objects in an image.

In addition, we use the Flickr30K dataset [61] to evaluate the model’s overall capability retention in multi-modal tasks. This dataset contains rich image–text pairs, with captions typically longer than those in VQA tasks, making it suitable for assessing the model’s comprehensive performance in multi-modal understanding and generation.

## B Baselines and Implementation Details

This section provides detailed descriptions of all continual learning baselines compared in this work, including their core ideas and optimization objectives. The baselines fall into three categories: **Regularization-based methods**, **Replay-based methods**, and **Parameter-Efficient Tuning (PET)**.

### B.1 Regularization-based Methods

#### EWC (Elastic Weight Consolidation).

EWC [24] estimates the importance of model parameters on previous tasks and applies a quadratic penalty to discourage significant updates on important parameters when learning new tasks. The overall objective for task  $t$  is :

$$\mathcal{L}_{\text{EWC}} = \mathcal{L}_t(\theta) + \lambda \sum_i F_i(\theta_i - \theta_i^*)^2, \quad (14)$$

where  $\theta^*$  denotes the parameter values after training the previous task,  $F_i$  is the Fisher information representing the importance of parameter  $i$  and  $\lambda$  controls the strength of the regularizer (set to 0.5 in our experiments).

#### GEM (Gradient Episodic Memory).

GEM [34] stores a small number of samples from previous tasks in an episodic memory and constrains gradient updates so that the loss on old tasks does not increase. During training on task  $t$ , GEM requires:

$$\langle g, g_k \rangle \geq 0, \quad \forall k < t, \quad (15)$$

where  $g$  is the gradient for the current task and  $g_k$  is the gradient computed on memory samples of task  $k$ . If this constraint is violated, i.e.,  $\langle g, g_k \rangle < 0$ , GEM solves the following quadratic program to obtain a projected gradient:

$$\begin{aligned} \min_{\tilde{g}} \quad & \| \tilde{g} - g \|^2 \\ \text{s.t.} \quad & \langle \tilde{g}, g_k \rangle \geq 0, \forall k < t. \end{aligned} \quad (16)$$

The projected gradient  $\tilde{g}$  is then used for parameter updates, ensuring no negative backward transfer.

#### LwF (Learning without Forgetting).

LwF [27] employs the predictions of the previous model as soft targets and introduces a distillation loss to constrain changes in the output distribution when learning new tasks:

$$\mathcal{L}_{\text{LwF}} = \mathcal{L}_t^{\text{CE}} + \alpha \cdot \mathcal{L}_{\text{KD}}, \quad (17)$$

where the distillation loss is:

$$\mathcal{L}_{\text{KD}} = - \sum_c p_c^{\text{old}}(T) \log p_c^{\text{new}}(T), \quad (18)$$

with softened probabilities computed using temperature  $T$ :

$$p_c(T) = \frac{\exp(z_c/T)}{\sum_j \exp(z_j/T)}. \quad (19)$$

Following common practice, we set  $\alpha = 0.5$  and  $T = 2$ .

## B.2 Replay-based Methods

Replay-based methods mitigate forgetting by reusing data from previous tasks when training new ones. We implement two representative replay strategies: **Replay (offline replay)**. After completing training on the current task, we perform an additional replay phase where we uniformly sample 1% of data from each previous task stored in a unified memory buffer. These real samples are used for extra gradient steps to reinforce previously learned knowledge.

**Replay-online (online replay)** During each training iteration of a new task, we dynamically sample approximately 1% of data from the same memory buffer and mix it with the current task’s mini-batch. This enables continuous rehearsal throughout training, improving stability in knowledge retention.

Both strategies rely solely on real data and share the same unified memory buffer. No synthetic data or auxiliary generative models are used. We also enforce uniform sampling across tasks to ensure fairness in the replay process.

## B.3 Parameter-Efficient Tuning (PET) Baselines

### SeqLoRA (Sequential Low-Rank Adaptation):

LoRA [21] introduces a trainable low-rank update to the weight matrix of attention layers. For a linear transformation  $h = Wx$ , LoRA reparameterizes the weight as:

$$W' = W + \Delta W, \quad \Delta W = BA, \quad (20)$$

where  $A \in \mathbb{R}^{r \times d}$ ,  $B \in \mathbb{R}^{d \times r}$ , and  $r = 8$  is the rank used in our experiments. A scaling factor  $\alpha$  is applied to stabilize optimization (set to 32):

$$\Delta W = \frac{\alpha}{r} BA, \quad (21)$$

We follow common practice and apply LoRA to the attention projection matrices q-proj and v-proj, with dropout set to 0.1. SeqLoRA shares the same LoRA parameters across all tasks, enabling a simple sequential PET baseline.

### O-LoRA (Orthogonal LoRA).

O-LoRA [56] prevents interference across tasks by enforcing orthogonality between the LoRA subspaces of different tasks. For LoRA parameters  $A_i$  and  $A_j$  of task  $i$  and  $j$ , the orthogonality regularizer is:

$$\mathcal{L}_{\text{ortho}} = \gamma \|A_i^\top A_j\|_F^2. \quad (22)$$

and the final test-time weight combines LoRA updates from all tasks:

$$W_{\text{test}} = W_0 + \sum_t B_t A_t. \quad (23)$$

LoRA hyperparameters follow the same settings as SeqLoRA, with orthogonality coefficient  $\gamma = 0.5$ .

### LayerNorm.

LayerNorm [65] updates only the LayerNorm parameters while keeping all other model weights frozen.

### MIGU (Magnitude-based Gradient Updating for continual learning).

MIGU [14] is an unstructured PET method that selectively updates parameters based on their output magnitude. For a linear layer with weight  $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$  and input  $x$ , the response of the  $i$ -th output channel is:

$$h_i = W_i \cdot x \quad (24)$$

and its importance score is the L1 norm:

$$n_i = \|h_i\|_1. \quad (25)$$

During training, only the top 0.1% parameters ranked by these scores are updated, while gradients on the remaining parameters are masked out, making the update pattern consistent with our proposed method.

## C Metrics

Let the performance score of the model on task  $i$  after learning task  $t$  be denoted as  $R_{t,i}$ . For classification tasks or question-answering tasks with single-token labels, the performance score is measured by **accuracy**; for other tasks, the performance score is the average of the **Rouge-L** and **BLEU** scores. We use the following metrics to assess the learning effects of different methods on sequential tasks:

### C.1 Overall Performance (OP)

Overall performance measures the average performance of the model across all learned tasks after training task  $t$ :

$$OP_t = \frac{1}{t} \sum_{i=1}^t R_{t,i} \quad (26)$$

### C.2 Backward Transfer (BWT)

Backward transfer quantifies how learning new tasks affects the performance on previously learned tasks:

$$BWT_t = \frac{1}{t} \sum_{i=1}^{t-1} (R_{t,i} - R_{i,i}) \quad (27)$$

where  $R_{i,i}$  is the performance on task  $i$  immediately after its training. Positive BWT indicates that new task learning improves old tasks, while negative BWT indicates forgetting. **Note:** OP and BWT are calculated only on sequential tasks to enable fair comparison with other baselines.

### C.3 Original Capability Retention

#### Language Models (Programming Ability)

For language models, we use the HumanEval dataset to

evaluate code generation ability. Pass@K is defined as:

$$\text{Pass@K} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(y_i \in \{\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \dots, \hat{y}_i^{(K)}\}), \quad (28)$$

where  $y_i$  is the ground-truth code for the  $i$ -th test case,  $\hat{y}_i^{(k)}$  is the  $k$ -th generated sample, and  $K = 1$  in this study.

#### Multimodal Models (Image Captioning Ability)

For multimodal models, we use the Flickr30K dataset, with performance measured as the average of Rouge-L and BLEU scores. BLEU-N is definded as:

$$\text{BLEU}_N = \text{BP} \cdot \exp \left( \frac{1}{N} \sum_{n=1}^N \log p_n \right), \quad (29)$$

$$\text{BP} = \begin{cases} 1, & c > r \\ e^{1-r/c}, & c \leq r \end{cases}$$

where  $c$  is the length of the generated sequence,  $r$  is the reference sequence length, and  $p_n$  is the n-gram precision. Rouge-L is definded as:

$$\text{ROUGE-L} = \frac{(1 + \beta^2) \cdot P \cdot R}{R + \beta^2 \cdot P},$$

$$P = \frac{\text{LCS}(X, Y)}{|X|}, \quad (30)$$

$$R = \frac{\text{LCS}(X, Y)}{|Y|},$$

$$\beta = 1$$

where  $X$  is the generated sequence,  $Y$  is the reference sequence, and  $\text{LCS}(X, Y)$  denotes the length of the longest common subsequence.

## D Extended Experimental Analysis

This section provides additional experimental details and analyses, including the distribution of key parameters in the PIECE-F and PIECE-S models, as well as the overlap of importance parameters across tasks and their corresponding correlation coefficients, and it presents the PIECE training logs of selected models, to help readers gain a more comprehensive understanding of how the proposed method operates.

### D.1 Distribution of Critical Parameters

Figure 8-12 illustrate the distribution of critical parameters across different models. We adopt a unified set of abbreviations for the major network modules within the model: the attention-related modules ( $q\_proj$ ,  $k\_proj$ ,  $v\_proj$ ,  $o\_proj$ ) are denoted as Q, K, V, and O, respectively; the feed-forward modules  $gate\_proj$ ,  $up\_proj$ , and  $down\_proj$  are denoted as G, U, and D; the normalization layers  $input\_layernorm$ ,

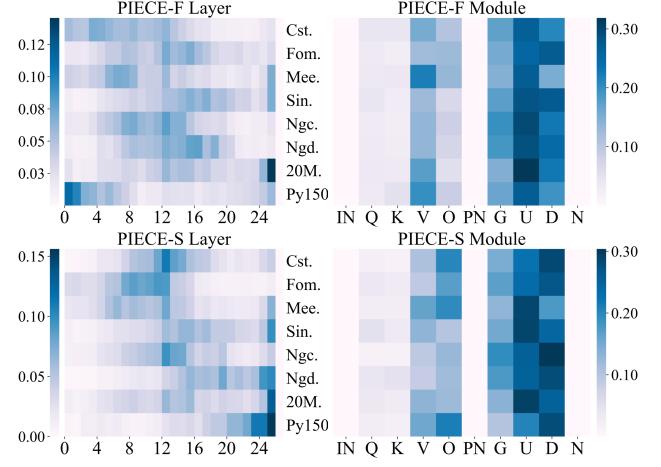


Figure 8. Distribution of critical parameters under Gemma2-2B.

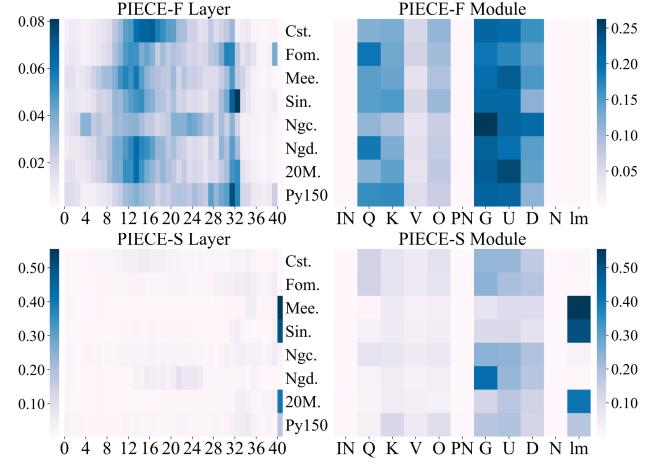


Figure 9. Distribution of critical parameters under Qwen3-14B ( $k = 0.1\%$ ).

post\_attention\_layernorm, and norm are denoted as IN, PN, and N; and the final output layer lm\_head is denoted as lm. These abbreviations are used consistently throughout the paper. To better observe the important parameters in large scale Qwen3-14B model, we expanded the proportion of selected important parameters from 0.1% to 1%, as shown in Figure 9 and 10.

Experimental results indicate that **architectural characteristics** and **task attributes** all potentially influence the distribution of important parameters within a network. First, **from the perspective of model architecture**, the distribution of important parameters differs significantly across architectures, while models within the same architecture exhibit a high degree of similarity. For example, despite notable differences in task types, the distributions of important parameters across layers and submodules in Qwen3-VL-4B and Qwen3-14B are largely consistent in Figure 10 and 11.

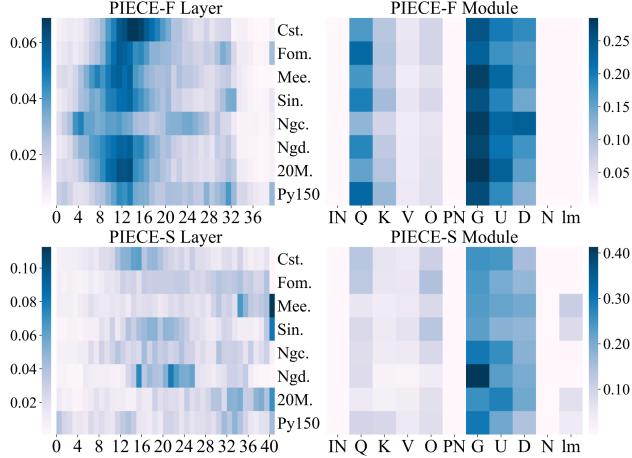


Figure 10. Distribution of critical parameters under Qwen3-14B ( $k = 1\%$ ).

Similarly, LLaVA-1.5-7B and Llama3-8B, which share the LLaMA architecture, show comparable patterns in Figure 3 and 12. Moreover, Gemma2-2B has the similar updating submodules with LLaMA architecture in Figure 8.

Second, **tasks attributes likely constitute another key factor influencing the distribution of important parameters**. Take the Qwen series as an example, although the 4B VLMs and 14B LMs share the same architecture, their preferences within feed-forward modules differ in Figure 10 and 11. The 4B VLMs tends to select higher-importance parameters in the up-proj (U) module, whereas the 14B LMs shows greater involvement in the gate-proj (G) module.

Although different models may vary in their objectives and intensity of parameter “protection” due to architectural and task differences, **all models consistently avoid updating certain parameters across layers and modules**. We posit that this parameter protection mechanism is a key reason why unstructured parameter-efficient fine-tuning achieves strong performance in continual learning.

Taken together, these observations suggest that model architecture and task attributes jointly influence the distribution of important parameters. While these hypotheses require further systematic investigation, they at least indicate that **evaluating important parameters through manual design or fixed heuristics is extremely challenging**, and they further highlight the practical significance of PIECE.

## D.2 Distribution Overlap and Correlation Across Tasks

Figures 13-16 illustrate the coverage of important parameters across tasks for various models. The results indicate that, in general, as model size increases, the proportion of parameters shared across tasks decreases, which aligns with intuition. Nevertheless, even for the smaller 2B model (Fig-

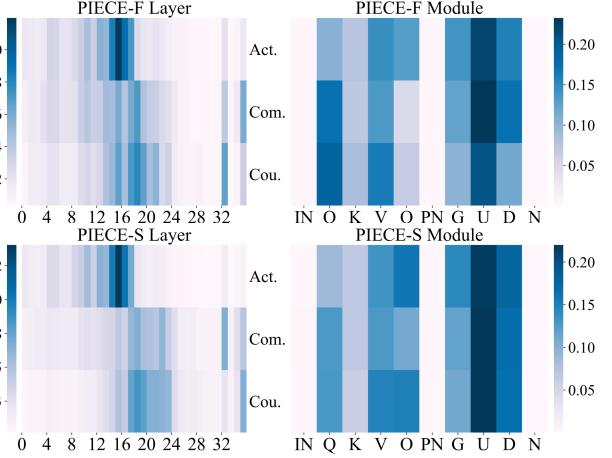


Figure 11. Distribution of critical parameters under Qwen3-VL-4B.

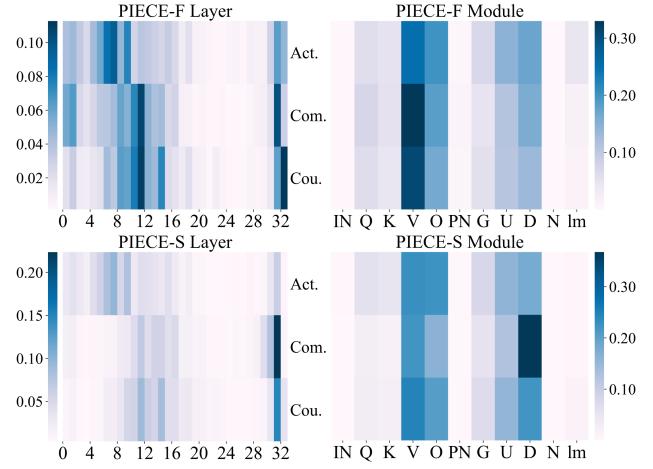


Figure 12. Distribution of critical parameters under LLaVA-1.5-7B.

	Gemma2-2B		Llama3-8B		Qwen3-14B	
	PIECE-F	PIECE-S	PIECE-F	PIECE-S	PIECE-F	PIECE-S
Pearson $r$	0.03	-0.06	0.01	-0.04	-0.39	-0.30
Pearson $p$	0.88	0.75	0.96	0.83	0.45	0.39
Spearman $\rho$	0.01	-0.17	0.06	-0.01	-0.42	-0.29
Spearman $p$	0.97	0.39	0.77	0.98	0.39	0.40

Table 4. Importance Parameter Correlations Across Tasks. Pearson  $r$  and Spearman  $\rho$  indicate correlation strength (closer to  $\pm 1$  = stronger correlation, near 0 = weak/no correlation);  $p$ -values indicate significance (closer to 0 = significant, near 1 = not significant).

ure 13), task-wise coverage remains relatively low. Across all model sizes, PIECE-S consistently exhibits lower coverage than PIECE-F, indicating greater task specificity in parameter allocation.

Moreover, although the absolute coverage values differ across models, the relative overlap patterns between tasks

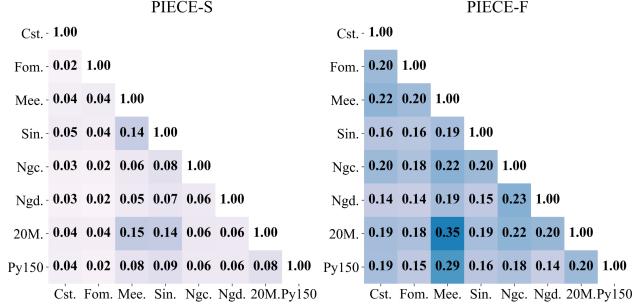


Figure 13. Parameter overlap across tasks under Gemma2-2B.

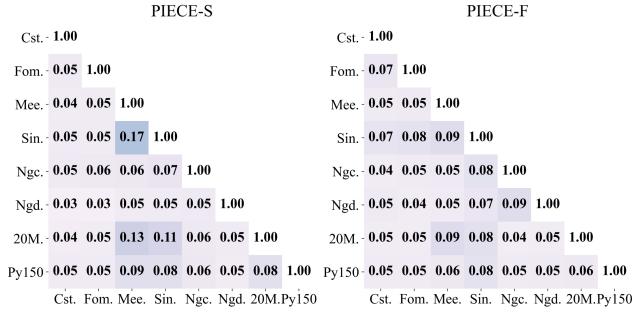


Figure 14. Parameter overlap across tasks under Qwen3-14B.

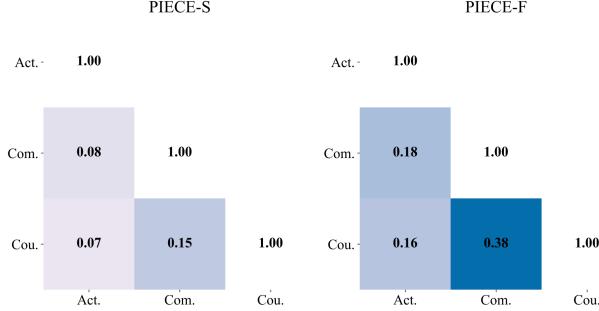


Figure 15. Parameter overlap across tasks under Qwen3-VL-4B.

remain largely consistent. For instance, if two tasks show higher parameter overlap than other task pairs in one model, this relative trend tends to persist in other models as well, suggesting that certain task relationships are reflected robustly in parameter sharing patterns, independent of model size.

Table 4 further quantifies the relationship between coverage and task forgetting, showing that task forgetting is weakly or very weakly correlated with the extent of parameter overlap. This indicates that while overlap exists between important parameters across tasks, it does not directly predict forgetting, highlighting the complex relationship between parameter sharing and continual learning performance.

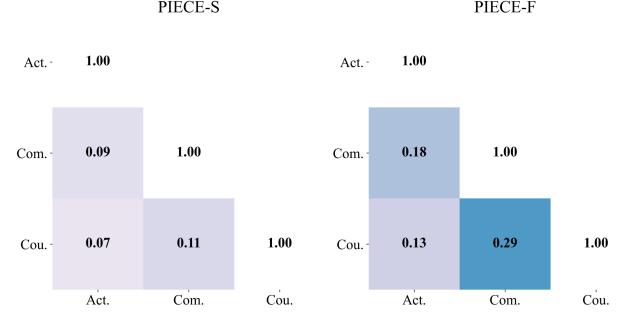


Figure 16. Parameter overlap across tasks under LLaVA-1.5-7B.

### D.3 Training Logs and Parameter Dynamics

We present several examples of PIECE training logs, as shown in Figures 17-19, which depict the loss trajectories of individual tasks throughout the continual learning process. To reduce computational overhead, we independently sampled 200 instances from each dataset that were not part of the training or test sets for loss evaluation. The dataset abbreviations on the horizontal axis indicate the task currently being trained.

The results suggest that, compared with SeqFT, PIECE effectively mitigates interference across tasks during training, allowing the model to progressively improve its overall performance in a continual learning setting. Notably, the mathematical reasoning tasks Ngc. and Ngd. appear to be more strongly influenced by other tasks, which may indicate that these tasks impose higher demands on the model’s integrated reasoning capabilities.

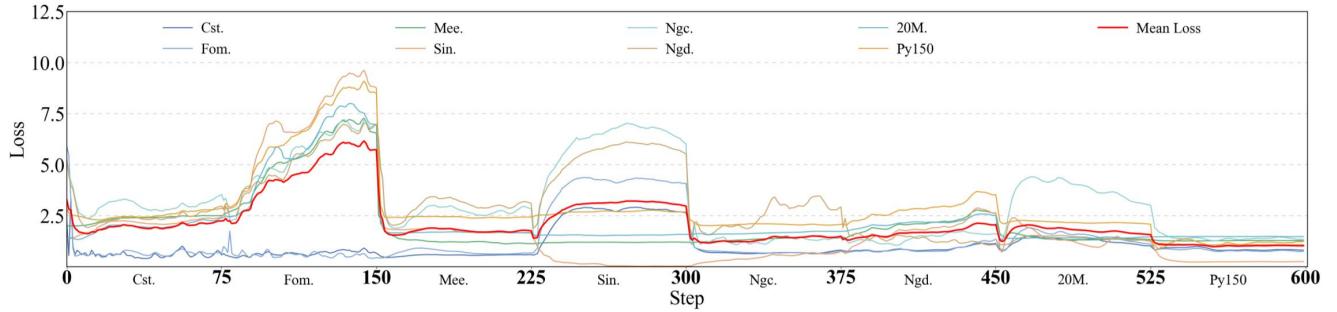


Figure 17. Multi-task Training Loss Curves for Llama3-8B (SeqFT).

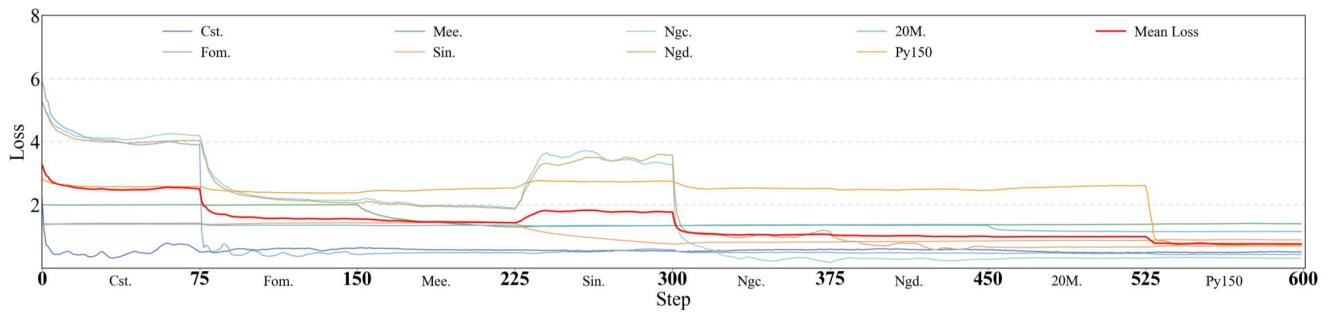


Figure 18. Multi-task Training Loss Curves for Llama3-8B (PIECE-F).

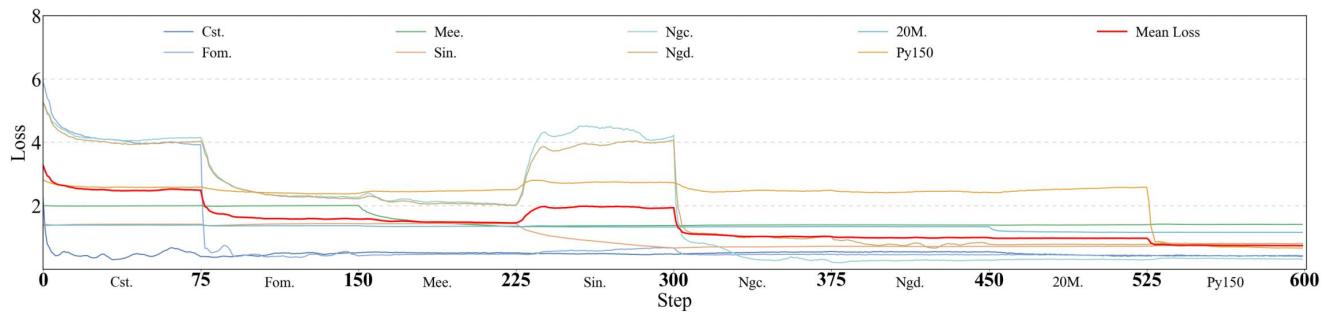


Figure 19. Multi-task Training Loss Curves for Llama3-8B (PIECE-S).