# Harmonizing the Past, Present, and Future: A Null-Space Constrained Region-Specific Method for Continual Learning in LLMs

**Jinhui Chen**[1,2], **Shizhu He**[1,2,3*], **Xingchang Yang**[1], **Huanxuan Liao**[1,2], **Yequan Wang**[3*], **Xiangwen Liao**[4], **Wenhao Teng**[5], **Kang Liu**[1,2], **Jun Zhao**[1,2]

[1] C2DL, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Beijing Academy of Artificial Intelligence, Beijing, China
[4] College of Computer and Data Science, Fuzhou University
[5] Department of Gastrointestinal Surgery, Fujian Provincial Cancer Hospital
chenjinhui2025@ia.ac.cn, shizhu.he@nlpr.ia.ac.cn

## Abstract

Enabling Large Language Models (LLMs) to evolve sustainably requires simultaneously preserving previously acquired knowledge (**Past**), effectively acquiring new task-specific skills (**Present**), and reserving sufficient parameter capacity for subsequent adaptation (**Future**). However, existing continual learning (CL) paradigms often prioritize immediate performance through dense updates, leading to catastrophic forgetting and rapid exhaustion of model capacity. To harmonize these conflicting demands, we draw inspiration from the brain's functional partitioning and propose the **Null-Space Constrained Parameter Region-Specific Method (PaRSP)**. PaRSP establishes a dynamic "Task-Region Mapping" that distinguishes between *specialized neurons* and *generalist neurons*. By precisely localizing a sparse "functional core" for each task, PaRSP restricts updates to specific regions via null-space orthogonality, preserving the vast majority of the network as an immutable "long-term memory bank." This induced sparsity not only enhances plasticity via targeted adaptation and minimizes interference to ensure stability, but also strategically reserves substantial capacity, securing sustainability for future evolution. Extensive experiments validate PaRSP's state-of-the-art performance, particularly on Standard CL and Long Sequence benchmarks, effectively harmonizing the stability-plasticity-sustainability trade-off. Code is available at https://github.com/JinhuiBot/PaRSP

## 1 Introduction

**Continual Learning (CL)** for Large Language Models (LLMs) (Dubey et al., 2024; Yang et al., 2025) has become increasingly essential for addressing evolving real-world scenarios, continuously updating domain knowledge, and accommodating diverse user demands (Zhou et al., 2024).

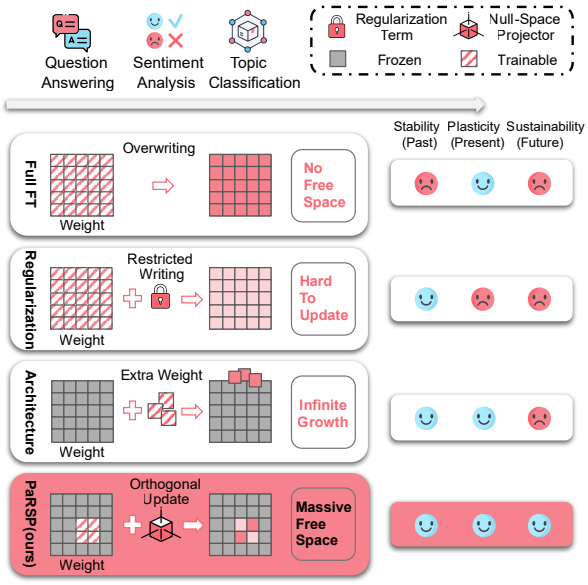


Figure 1: Comparison of Continual Learning paradigms across three key capabilities: **Past** (Stability), **Present** (Plasticity), and **Future** (Sustainability). (1) **Full Fine-Tuning** prioritizes the present, sacrificing past stability and saturating future capacity. (2) **Regularization** constrains updates to protect the past, hindering present and future plasticity. (3) **Architecture-based methods** incur unbounded parameter growth, compromising future sustainability. (4) **Our PaRSP** harmonizes the triad via sparse activation (plasticity) and null-space orthogonality (stability), reserving capacity for future sustainability.

An ideal CL method should not only handle current streaming tasks efficiently but also maintain cognitive coherence and optimal resource allocation over long-term knowledge evolution, thereby facilitating a transition from "static models" to "dynamically evolving intelligence." Specifically, such a method must support three critical capabilities: 1) **Stability (Preserving the Past)**: The ability to preserve and consolidate previously acquired knowledge, effectively mitigating catastrophic forgetting (McCloskey and Cohen, 1989); 2) **Plasticity (Learning the Present)**: The capacity to rapidly adapt to new

* Corresponding authors.



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27013–27033
July 2-7, 2026 

tasks and acquire novel knowledge efficiently (Dohare et al., 2024); and 3) **Sustainability (Reserving for the Future)**: The capacity for long-term adaptation, preserving sufficient parameter capacity and learning effectiveness for future challenges.

Although recent studies have explored a wide range of strategies to improve continual learning in large language models, achieving a unified balance among these objectives remains unresolved. As illustrated in Figure 1: (1) **Full Fine-Tuning** (Ouyang et al., 2022) prioritizes the present via global updates. Despite immediate gains, this triggers catastrophic forgetting of the past and prematurely saturates capacity, compromising future expansion. (2) **Regularization-based** methods (Kirkpatrick et al., 2017; Du et al., 2024) preserve the past via rigid constraints or selective weight freezing. However, this "defensive" rigidity severely restricts plasticity, hindering effective adaptation to the present and future. (3) **Architecture-based** methods (Wang et al., 2023a; Razdaibiedina et al., 2023) mitigate interference by adding dedicated modules. Yet, the resulting linear parameter growth incurs poor efficiency, rendering them unsustainable for the future. Therefore, **the greatest challenge of continual learning in LLMs is to balance immediate plasticity for current tasks with long-term stability of past knowledge, while preserving capacity for future adaptation**.

Notably, the human brain demonstrates remarkable resilience in continual learning despite finite biological resources. This capability stems from a sophisticated storage mechanism where memory is encoded within sparse, task-specific clusters known as *memory engrams* (Josselyn and Tonegawa, 2020), rather than in a homogeneous medium. These engrams are dynamic, utilizing *systems consolidation* to reorganize circuitry for long-term retention without disrupting new acquisition (Ko et al., 2025). Crucially, biological circuits employ a *functional partitioning* mechanism: while certain neurons are highly specialized for specific tasks, others serve as shared bridges for general cognition (Sporns and Betzel, 2016). This duality suggests that effective knowledge representation implies inherent modularity and sparsity (Olshausen and Field, 2004). By emulating these strategies, LLMs can effectively balance task-specific adaptation with global cognitive preservation, achieving a synergy between stability, plasticity, and sustainability.

Inspired by these insights, we propose the **Null-Space Constrained Parameter Region-Specific Method (PaRSP)** for LLMs. Our core motivation is to establish a dynamic "Task-Region Mapping" that explicitly distinguishes between **specialized neurons** (prioritizing plasticity) and **generalist neurons** (requiring stability). Through precisely localizing the "functional core" of each new task, PaRSP identifies the optimal parameter subspace for efficient adaptation to ensure high **Plasticity**, while simultaneously confining gradient updates to sparse regions to sequester the vast majority of the network as an immutable memory repository. Enforcing null-space orthogonality within these active regions further ensures that updates remain algebraically non-destructive to established representations. This synergistic design not only mitigates cross-task interference to bolster **Stability** but also strategically conserves substantial parameter capacity, thereby securing long-term **Sustainability** for future adaptation.

Specifically, PaRSP operates via a synergistic dual-stage mechanism: *Task-Specific Parameter Region Activation* followed by *Orthogonal Update via Null Space Projection*. In the first stage, we utilize neuron-level attribution (Yu and Ananiadou, 2024) to pinpoint the salient "functional core" most relevant to the current task. By topologically isolating these specialized neuron clusters, we maximize present plasticity while sequestering parameter capacity for future sustainability. In the second stage, we leverage **Null Space Projection theory** (Saha et al., 2021; Wang et al., 2023a) to mathematically rectify weight updates $\Delta\mathbf{W}$ within the identified active regions. Constraining these updates to the null space of prior activations ensures representation invariance, thereby rigorously safeguarding past stability.

We conducted extensive evaluations on multiple benchmarks, including standard CL benchmarks (Zhang et al., 2015), long-sequence benchmarks (Razdaibiedina et al., 2023), and TRACE (Wang et al., 2023b), using Llama-3.1-8B (Dubey et al., 2024) and Qwen-2.5-7B (Yang et al., 2025). Results demonstrate that PaRSP establishes new state-of-the-art performance on the Standard CL and Long Sequence benchmarks, offering a robust paradigm that effectively harmonizes the past, present, and future of LLM evolution.

Our main contributions are summarized as follows:

- We propose **PaRSP**, a novel bio-inspired CL method that, to the best of our knowledge, is the first to explicitly harmonize preserving past knowledge, learning current tasks, and reserving capacity for future adaptation.

- We integrate **Null Space Projection theory** into task-specific region updates, constructing a dual-protection mechanism that theoretically guarantees resistance to catastrophic forgetting.

- We introduce an optimized **Dual SVD strategy** and *implicit projection* mechanism, which shifts computational complexity to a linear scale. This renders PaRSP production-ready for massive LLMs with negligible precomputation overhead.

- Extensive experiments across three diverse benchmark categories show PaRSP achieves highly competitive performance, establishing new SOTA results on the Standard CL and Long Sequence benchmarks while remaining robust on complex reasoning tasks.

## 2 Related Work

### 2.1 Continual Learning For LLMs

Continual Learning (CL) aims to acquire knowledge from task streams without erasing previously learned information. Classical approaches primarily fall into three categories: *Rehearsal-based methods* mitigate forgetting by replaying history, either via synthesized pseudo-samples (Shin et al., 2017; Sun et al., 2020) or buffered real data employed for gradient constraints (Lopez-Paz and Ranzato, 2017; Chaudhry et al., 2019) and coreset optimization (Tiwari et al., 2022; Wang et al., 2024; He et al., 2024); *Regularization-based methods* (Kirkpatrick et al., 2017; Du et al., 2024) impose constraints on critical weights to prevent drastic updates; and *Architecture-based methods* (Wang et al., 2023a; Razdaibiedina et al., 2023) dynamically expand model capacity or isolate task-specific parameters. However, these paradigms suffer from privacy risks, restricted plasticity, or unbounded expansion. In contrast, **PaRSP** optimizes a fixed parameter space via strategic partitioning—allocating mutable regions for adaptation and immutable cores for memory—thereby circumventing replay or architectural growth while ensuring sustainable evolution.

### 2.2 Neuron-Level Interpretability and Optimization

Research on neuron-level interpretability has evolved from passive analysis to active optimization, aiming to elucidate and manipulate how LLMs encode knowledge. Early analytic efforts primarily utilized gradient-based attribution (Sundararajan et al., 2017) to identify "Knowledge Neurons" (Dai et al., 2022) or conceptualized Transformer layers as *Key-Value Memories* (Geva et al., 2021). Crucially, these insights have been translated into targeted **optimization strategies**. For instance, causal tracing methods have enabled precise *Model Editing* (Meng et al., 2022, 2023), allowing for the surgical correction of factual errors by modifying specific MLP weights. Similarly, Wang et al. (2022) identified "Skill Neurons" to guide parameter-efficient tuning, demonstrating that optimizing only a tiny subset of task-specific neurons yields competitive performance. However, given the prohibitive costs of dynamic intervention, we instead leverage efficient static attribution (Yu and Ananiadou, 2024; Tan et al., 2025). This allows PaRSP to rapidly localize task-specific "functional cores," bypassing the scalability bottlenecks of repeated forward-backward passes.

## 3 Problem Formulation

Continual Learning (CL) enables models to sequentially acquire new knowledge without catastrophically forgetting prior information.

**Formal Definition.** Let $\mathcal{M}_\theta$ denote an LLM parameterized by $\theta$. The model encounters a task stream $\mathcal{T} = \{T_1, \ldots, T_N\}$, where each task $T_t$ is associated with a data distribution $\mathcal{D}_t$ and a loss function $\mathcal{L}_t$. The objective is to minimize the expected loss on the current task $T_t$ (updating parameters to $\theta_t$) subject to the constraint that performance on all previous tasks $j < t$ remains non-degrading:

$$\min_{\theta_t} \quad \mathbb{E}_{(x,y)\sim\mathcal{D}_t}[\mathcal{L}_t(\mathcal{M}_{\theta_t}(x), y)] \tag{1}$$

$$\text{s.t.} \quad \forall j < t, \quad \mathbb{E}_{(x,y)\sim\mathcal{D}_j}[\mathcal{L}_j(\mathcal{M}_{\theta_t}(x), y)] \leq \mathbb{E}_{(x,y)\sim\mathcal{D}_j}[\mathcal{L}_j(\mathcal{M}_{\theta_j}(x), y)] \tag{2}$$

where $\theta_j$ represents the model's parameters after completing task $T_j$. The constraint in Equation (2) precisely defines the mathematical desideratum of "non-forgetting."

**Study Setup.** We adopt a rigorous setting characterized by: (1) **Heterogeneous Task Sequences**,

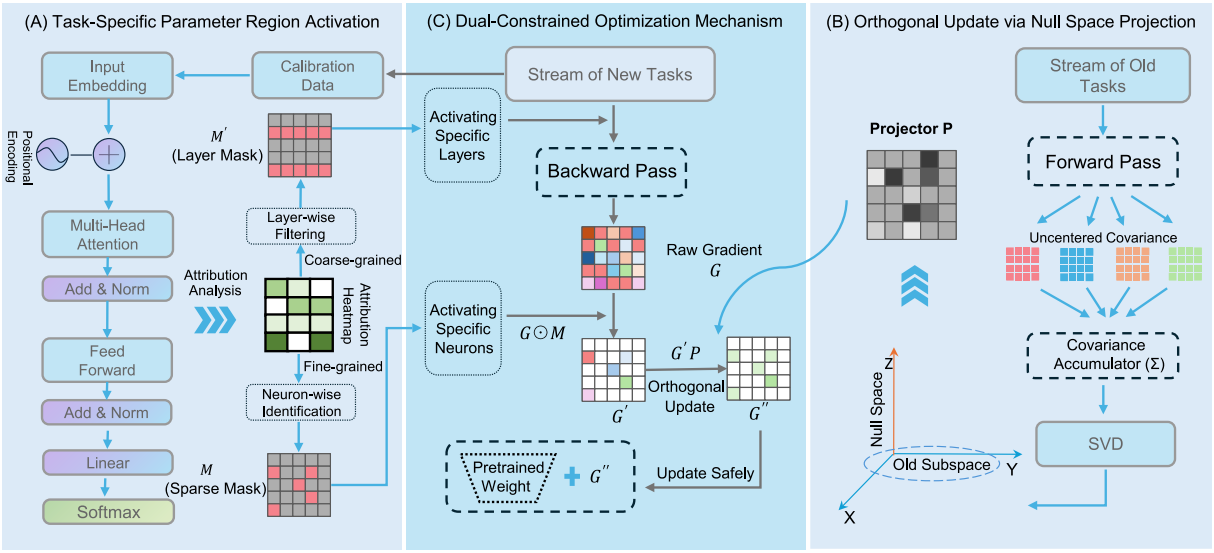


Figure 2: **Overview of the Null-Space Constrained Parameter Region-Specific Method (PaRSP).** The method operates in a synergistic dual-stage process and coupled with a unified optimization process: **(A) Task-Specific Parameter Region Activation:** Static attribution is employed to generate a binary mask $\mathbf{M}$ that topologically isolates task-specific neurons (highlighted in pink), while freezing the remaining parameters. **(B) Null-Space Projector Construction:** Uncentered covariance matrices accumulated from prior tasks are used to construct a projector $\mathbf{P}$ via singular value decomposition (SVD). From a geometric perspective, $\mathbf{P}$ constrains updates to lie in the null space, ensuring orthogonality to previously learned knowledge. **(C) Dual-Constrained Optimization:** During training, the gradient updates are jointly regulated by the spatial mask $\mathbf{M}$ and the directional projector $\mathbf{P}$, ensuring that the resulting parameter update $\Delta\mathbf{W}$ is both sparse and non-interfering.

encompassing diverse tasks that require implicit format distinction; and (2) **Task-Agnostic Evaluation**, where the model must infer objectives directly from the input context without explicit task identifiers.

## 4 Methodology

### 4.1 Motivation and Overview

Our methodology operationalizes the biological principle of *functional partitioning* highlighted in the Introduction. Just as the brain encodes memory in sparse, task-specific *engrams* to optimize efficiency and reduce interference, we aim to implement a similar dual-protection mechanism in LLMs. We translate this biological insight into a method that imposes *topological isolation* (mimicking sparse allocation) and *geometric constraints* (mimicking synaptic stability). Specifically, by confining updates to a sparse "functional core" and projecting gradients onto the *null space* of prior activations, we ensure that the model's evolution is "lossless"—granting the freedom to adapt to new tasks while mathematically guaranteeing the invariance of historical knowledge.

Based on this philosophy, we propose the **Null-Space Constrained Parameter Region-Specific Method (PaRSP)**. As illustrated in Figure 2, PaRSP harmonizes the *Past*, *Present*, and *Future* via two synergistic stages:

**(1) Task-Specific Parameter Region Activation.** Facing a new task, we employ a lightweight attribution-based localization mechanism to identify the most semantically relevant neuron clusters. This process mimics the brain's sparse activation strategy, allowing us to freeze the vast majority of the network and activate only a specific subset. This topological isolation maximizes plasticity for the present while strategically reserving massive parameter capacity for the future (Sustainability).

**(2) Orthogonal Update via Null Space Projection.** Within the selected active regions, we enforce a null-space constrained optimization. We derive the null space from the input activations of prior tasks and project new task gradients onto it. Geometrically, this ensures the update trajectory is strictly orthogonal to the feature subspace of old knowledge. This design enables the superposition of capabilities within the same physical region, ensuring that acquiring the Present task does not compromise the stability of the past.

### 4.2 Task-Specific Parameter Region Activation

To operationalize the biological principle of functional partitioning, we aim to identify and activate only the subset of parameters responsible for the target task, while freezing the remainder. However, locating these "task-specific regions" within the vast parameter space of LLMs is non-trivial. Drawing inspiration from the *Neuron-Level Knowledge Attribution* framework (Yu and Ananiadou, 2024), we formulate a strategy based on information salience.

**Definition of Attribution Score.** We define the importance of a model component $v$ (e.g., a specific layer output or a neuron's contribution) by measuring its impact on the correct prediction of the target token $y$. Leveraging the *Logit Lens* approach, we project the latent state produced by $v$ onto the vocabulary space and compute the **Log-Probability Gain**:

$$\text{Imp}(v) = \frac{1}{|\mathcal{D}_k|} \sum_{(x,y)\in\mathcal{D}_k} \Big[ \log P(y \mid h_{in} + \Delta h_v) - \log P(y \mid h_{in}) \Big] \tag{3}$$

where $h_{in}$ denotes the input representation (residual stream) accumulated before component $v$, $\Delta h_v$ represents the specific output vector contributed by $v$, and $\mathcal{D}_k$ is the calibration dataset for the current task $k$.

**Hierarchical Localization.** To balance computational efficiency with localization precision, we employ a coarse-to-fine hierarchical selection process to identify key components based on the scores defined in Equation (3). The detailed algorithmic procedures for layer-wise filtering and neuron-wise identification are provided in Appendix B.

This process yields a binary mask $\mathbf{M}^{(k)}$, which topologically isolates the "functional sub-network" for task $k$. By strict localization, we ensure that the model adaptation is confined to relevant regions, thereby maximizing plasticity for the present while reserving the majority of the parameter space to ensure sustainability for the future.

### 4.3 Orthogonal Update via Null Space Projection

While the specific area activation described in Section 4.2 provides a macroscopic isolation, a microscopic challenge remains. Neural circuits are not perfectly disentangled; within activated task-specific regions, neurons often exhibit a dual nature. We categorize them into two types: *Specialized Neurons*, which exclusively serve the current task, and *Generalist Neurons*, which act as cognitive bridges shared across multiple tasks. To update specialized neurons without disrupting the synaptic stability of generalist neurons, we introduce a strict geometric constraint: the **Null Space Projection**.

**Mathematical Formulation of the Null Space.** Our objective is to ensure that the parameter update $\Delta\mathbf{W}$ for the current task does not alter the model's response to previous tasks. Let $\mathbf{X}_{old} \in \mathbb{R}^{d\times N}$ denote the input activation matrix accumulated from previous tasks. To preserve existing knowledge, the updated weights $\mathbf{W}' = \mathbf{W} + \Delta\mathbf{W}$ must satisfy:

$$(\mathbf{W} + \Delta\mathbf{W})\mathbf{X}_{old} = \mathbf{W}\mathbf{X}_{old} \implies \Delta\mathbf{W}\mathbf{X}_{old} = 0 \tag{4}$$

This implies that the gradient update $\Delta\mathbf{W}$ must lie in the *Null Space* of the previous input representations, denoted as $\mathcal{N}(\mathbf{X}_{old})$. Geometrically, the learning trajectory of the new task must be orthogonal to the feature subspace spanned by old tasks, providing a mathematical guarantee for the stability of the past.

**Uncentered Covariance-Based Projector Construction.** Directly storing $\mathbf{X}_{old}$ is impractical. We leverage a fundamental linear algebra property: the null space of a matrix is identical to the null space of its *uncentered covariance matrix*, i.e., $\mathcal{N}(\mathbf{X}_{old}) \equiv \mathcal{N}(\mathbf{X}_{old}\mathbf{X}_{old}^T)$. This allows us to maintain only the recursive uncentered covariance matrix $\mathbf{\Sigma} = \sum_i \mathbf{x}_i\mathbf{x}_i^T$. We construct a projection matrix $\mathbf{P} = \mathbf{I} - \mathbf{U}_{core}\mathbf{U}_{core}^T$ via Singular Value Decomposition (SVD) on $\mathbf{\Sigma}$, where $\mathbf{U}_{core}$ spans the feature space of old tasks. During backpropagation, we project the raw gradient $\mathbf{g}$ onto the null space: $\hat{\mathbf{g}} = \mathbf{g}\mathbf{P}$. Consequently, parameter updates operate exclusively in directions that are "invisible" to the generalist neurons' historical function.

### 4.4 Dual-Constrained Optimization Mechanism

By integrating *Task-Specific Parameter Region Activation* (Section 4.2) with *Orthogonal Update via Null Space Projection* (Section 4.3), we propose a **Dual-Constrained Optimization** paradigm. Let $\mathcal{L}_{\text{task}}$ be the current loss and $\mathbf{G}_l = \nabla_{\mathbf{W}_l}\mathcal{L}_{\text{task}}$ be the raw gradient. To harmonize stability, plasticity, and sustainability, we simultaneously impose

| | Methods | Standard CL (SC) | | | Long Sequence (LS) | | | TRACE | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | FP ↑ | AP ↑ | Forget ↓ | FP ↑ | AP ↑ | Forget ↓ | FP ↑ | AP ↑ | Forget ↓ |
| Llama-3.1-8B | SeqLoRA | $79.6_{\pm.62}$ | $80.8_{\pm.49}$ | 5.8 | $74.8_{\pm.58}$ | $83.8_{\pm.52}$ | 9.0 | $65.1_{\pm.69}$ | $82.4_{\pm.54}$ | 17.3 |
| | LoRAReplay | $80.3_{\pm.71}$ | $80.9_{\pm.56}$ | 0.6 | $82.0_{\pm.69}$ | $85.0_{\pm.75}$ | 3.0 | **78.7**$_{\pm.83}$ | **85.7**$_{\pm.68}$ | 7.0 |
| | O-LoRA (Wang et al., 2023a) | $72.3_{\pm.86}$ | $73.9_{\pm.68}$ | 1.6 | $71.4_{\pm.64}$ | $74.8_{\pm.64}$ | 3.7 | $36.7_{\pm.57}$ | $50.1_{\pm.37}$ | 13.4 |
| | DATA (Liao et al., 2025) | $80.9_{\pm.40}$ | $80.6_{\pm.35}$ | -0.3 | $80.0_{\pm.53}$ | $82.3_{\pm.48}$ | 2.3 | $72.7_{\pm.94}$ | $80.4_{\pm.88}$ | 7.7 |
| | DATA + Replay | $80.8_{\pm.33}$ | $80.4_{\pm.47}$ | **-0.4** | $82.2_{\pm.66}$ | $82.6_{\pm.77}$ | 0.4 | $77.6_{\pm.37}$ | $81.0_{\pm.72}$ | **3.4** |
| | **PaRSP (Ours)** | **86.3**$_{\pm.52}$ | **86.0**$_{\pm.34}$ | -0.3 | **88.0**$_{\pm.67}$ | **88.2**$_{\pm.63}$ | **0.2** | $73.8_{\pm.85}$ | $80.1_{\pm.86}$ | 6.3 |

Table 1: Main results on Standard CL (SC), Long Sequence (LS), and TRACE benchmarks using the Llama-3.1-8B backbone. We report the mean and standard deviation over three independent runs. ↑: Higher is better; ↓: Lower is better. **Bold** denotes the best performance.

two constraints on $\mathbf{G}_l$: (1) a **Spatial Constraint** ($\mathbf{G}_{spatial} = \mathbf{G}_l \odot \mathbf{M}_l$), which enforces structural sparsity to maximize plasticity for the present while reserving capacity for the future; and (2) a **Directional Constraint** ($\mathbf{G}_{dual} = \mathbf{G}_{spatial}\mathbf{P}_l$), which projects gradients onto the null space of the historical uncentered covariance matrix to mathematically guarantee the stability of the past. To implement this efficiently without forward-pass overhead, we utilize a gradient rectification strategy via backward hooks (details in Appendix C).

Consequently, the final parameter update rule is formalized as:

$$\mathbf{W}_l^{(t+1)} \leftarrow \mathbf{W}_l^{(t)} - \eta \cdot [(\nabla_{\mathbf{W}_l} \mathcal{L}_{\text{task}} \odot \mathbf{M}_l)\mathbf{P}_l] \quad (5)$$

where $\eta$ is the learning rate. This formula encapsulates our core philosophy: *evolve only where necessary (via Mask), and move only where safe (via Projector).*

### 4.5 Efficiency and Scalability Analysis

While null-space projection typically entails cubic complexity associated with SVD, PaRSP is engineered to be highly tractable and scalable through a hierarchical efficiency design. We address computational and memory bottlenecks from two perspectives: foundational optimizations for general applicability, and advanced strategies for massive LLMs.

**Foundational Efficiency: Sparsity and Recursive Aggregation.** Fundamentally, our method mitigates computational overhead through *static attribution* and *induced sparsity*. By restricting expensive SVD operations strictly to a small, task-specific parameter subset ($\sim 5\% - 17\%$), the majority of the network is computationally skipped during projector construction. Furthermore, we address the linear memory growth issue ($\mathcal{O}(T \cdot d^2)$) inherent in standard rehearsal or projection methods via a **recursive uncentered covariance aggregation** strategy. By recursively accumulating historical statistics into a single dense matrix, we ensure a constant memory complexity of $\mathcal{O}(d^2)$ regardless of the task sequence length $T$. We provide a detailed theoretical analysis of these foundational properties and memory scalability in Appendix A.

**Advanced Scalability for Massive LLMs: Dual SVD & Implicit Projection.** While the $\mathcal{O}(d^2)$ memory footprint and sparse $\mathcal{O}(d^3)$ computation are tractable for standard models, they become prohibitive for massive LLMs. For instance, in Llama-3.3-70B, intermediate dimensions reach $d \approx 28,672$; explicitly constructing and decomposing a $d \times d$ covariance matrix incurs $\sim 2.3 \times 10^{13}$ FLOPs and requires $\sim$ 3.3 GB of memory per layer, inevitably triggering Out-Of-Memory (OOM) failures on standard GPUs. To render PaRSP production-ready for industrial-scale models, we mathematically restructure the projection mechanism by leveraging the extremely low-rank nature of our calibration set ($N = 256 \ll d$). We introduce a three-stage optimization pipeline:

**(1) Compute Optimization via Dual SVD.** Instead of $O(d^3)$ decomposition of $\mathbf{X}\mathbf{X}^T$, we compute the Gram matrix $\mathbf{K} = \mathbf{X}^T\mathbf{X} \in \mathbb{R}^{N \times N}$ in $O(dN^2)$. Based on duality, we perform SVD on the tiny matrix $\mathbf{K} = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T$ and recover the left singular vectors spanning the feature subspace via:

$$\mathbf{U} = \mathbf{X}\mathbf{V}\mathbf{\Sigma}^{-1} \quad (6)$$

This shifts complexity to a linear scale $\mathcal{O}(dN^2 + N^3)$, accelerating pre-computation from minutes to milliseconds.

**(2) Storage Optimization via Incremental Basis.** To avoid $\mathcal{O}(d^2)$ persistent storage, we maintain a compact basis $\mathbf{U}_{hist} \in \mathbb{R}^{d \times r}$ ($r \approx 128$). New tasks update the global basis recursively

via Dual SVD on the compressed matrix $\mathbf{M} = [\mathbf{U}_{hist}\mathbf{\Sigma}_{hist}, \mathbf{X}_{new}] \in \mathbb{R}^{d\times(r+N)}$, ensuring constant $\mathcal{O}(dr)$ memory complexity.

**(3) Memory-Efficient Implicit Projection.** We bypass instantiating the dense projector $\mathbf{P} = \mathbf{I} - \mathbf{U}\mathbf{U}^T \in \mathbb{R}^{d\times d}$ during backpropagation. Instead, a PyTorch backward hook utilizes matrix associativity to perform *Implicit Projection* on the raw gradient $\mathbf{g} \in \mathbb{R}^{1\times d}$:

$$\text{Proj}(\mathbf{g}) = \mathbf{g}\mathbf{P} = \mathbf{g} - (\mathbf{g}\mathbf{U}_{hist})\mathbf{U}_{hist}^T \quad (7)$$

This reduces peak memory from $\mathcal{O}(d^2)$ to $\mathcal{O}(dr)$, enabling PaRSP to scale indefinitely to industrial-scale models.

## 5 Experiments

### 5.1 Experimental Setup

To comprehensively evaluate the efficacy and robustness of PaRSP, we conduct experiments across three distinct continual learning benchmarks characterized by varying task types and sequence lengths. To ensure a direct and fair comparison with the current state-of-the-art, our experimental setup, including the reported scores for all baseline models, is meticulously aligned with the protocols established in **DATA** (Liao et al., 2025).

**Datasets.** We conduct experiments on three benchmarks: (1) **Standard CL Benchmark (SC)** (Zhang et al., 2015): Following Wang et al. (2023a), we select four datasets—AG News, Amazon Reviews, DBpedia, and Yahoo Answers—and shuffle them to construct three distinct task sequences (Orders 1–3). (2) **Long Sequence Benchmark (LS)** (Razdaibiedina et al., 2023): This extends the evaluation to 15 datasets, including classification tasks, GLUE, SuperGLUE, and IMDB. Adhering to the standard protocol, we generate three unique task sequences (Orders 4–6). (3) **TRACE** (Wang et al., 2023b): Designed for LLMs, it comprises 8 diverse datasets covering multi-choice QA, multilingual capabilities, code generation, and mathematical reasoning. Comprehensive details regarding dataset statistics, specific task sequences, and evaluation metric definitions are provided in Appendix D.1.

**Evaluation Metrics.** We quantify plasticity and stability using three standard metrics. Let $a_{i,j}$ denote the test performance on task $i$ after training on task $j$. We report: (1) **Average Performance (AP)**, which measures plasticity by averaging immediate task accuracies (AP $= \frac{1}{N}\sum_{j=1}^{N} a_{T_j,j}$); (2) **Final Performance (FP)**, which evaluates global retention after the sequence concludes (FP $= \frac{1}{N}\sum_{j=1}^{N} a_{T_j,N}$); and (3) **Forget**, which quantifies stability loss as the divergence between peak and final performance (Forget $=$ AP $-$ FP) (Wu et al., 2022; Jiang et al., 2025). Higher AP/FP and lower Forget indicate superior performance.

| Methods | MMLU | BBH | GSM8K | AGIEval | FP |
|---|---|---|---|---|---|
| Zero-Shot | 65.65 | 62.12 | 56.33 | 17.72 | - |
| SeqLoRA | 63.58 | 11.90 | 0.00 | **20.60** | 79.92 |
| LoRAReplay | 60.24 | 5.99 | 1.82 | 10.69 | 80.13 |
| O-LoRA | 62.79 | 6.31 | 1.56 | 13.87 | 71.83 |
| DATA | 64.47 | 10.42 | 3.63 | 16.94 | 81.39 |
| **PaRSP (Ours)** | **66.72** | **56.74** | **56.03** | 20.03 | **86.30** |

Table 2: Task generalization comparisons on unseen tasks using the Llama-3.1-8B backbone after training in Order 1. **Bold** indicates the best performance.

**Baselines.** To evaluate the effectiveness of PaRSP, we benchmark it against five representative methods: **SeqLoRA** sequentially trains standard LoRA modules without explicit anti-forgetting mechanisms; **LoRAReplay** augments SeqLoRA with a 2% experience replay buffer to mitigate forgetting; **O-LoRA** (Wang et al., 2023a) minimizes inter-task interference by enforcing orthogonality between task-specific subspaces; and **DATA** (Liao et al., 2025) optimizes the stability-plasticity trade-off by explicitly decoupling knowledge into high-rank and low-rank adapters via a decomposed attention mechanism.

**Implementation.** Experiments employ **Llama-3.1-8B** (Dubey et al., 2024) and **Qwen-2.5-7B** (Yang et al., 2025) on a single NVIDIA A100 GPU (80GB). We report the average results over three independent runs; detailed hyperparameters and configurations are provided in Appendix D.3.

### 5.2 Main Results

We evaluated PaRSP on three CL benchmarks using Llama-3.1-8B and Qwen-2.5-7B. Table 1 presents the aggregated results. Adhering to Wang et al. (2023a), we report averages across three independent runs with permuted task orders for SC and LS (detailed breakdowns provided in Appendix E.1).

**(1) Superior Stability: Preserving the Past.** PaRSP significantly bolsters the model's capacity to retain historical knowledge without external augmentation. Traditional methods like SeqLoRA

suffer from severe catastrophic forgetting across all benchmarks. While baselines such as LoRAReplay and DATA (+Replay) achieve competitive stability, they rely on auxiliary *replay mechanisms*, which often incur additional computational or storage overhead. In stark contrast, PaRSP achieves state-of-the-art (SOTA) performance in a strictly *rehearsal-free* manner. For instance, on the Standard CL (SC) and Long Sequence (LS) benchmarks, PaRSP achieves Final Performance (FP) scores of **86.3%** and **88.0%**, respectively, surpassing the previous SOTA method DATA by substantial margins of **5.5%** and **5.8%**. On the more challenging TRACE benchmark, PaRSP maintains an impressive FP of **73.8%** with a minimal Forget rate of **6.3%**. Notably, these results—achieved by updating approximately **17%** of neurons—outperform all rehearsal-free baselines and rival complex replay-based methods, validating the efficacy of our brain-inspired topological isolation in ensuring rigorous stability for the past.

**(2) Sustainable Plasticity: Present and Future.** Crucially, PaRSP resolves the stability-plasticity dilemma while securing long-term sustainability. While methods like DATA demonstrate reasonable stability, their plasticity is often constrained; even with replay, DATA's Average Performance (AP) on LS plateaus at 82.2%, indicating that its stability comes at the cost of suppressed new learning. Conversely, PaRSP maintains exceptional plasticity. On the LS benchmark, it attains an AP of **88.2%**, significantly outperforming DATA and LoRAReplay. Similarly, on TRACE, PaRSP exhibits plasticity comparable to DATA (+Replay). Most importantly, these gains are realized via a **non-greedy resource allocation** strategy. By activating only a sparse functional core ($\sim 17\%$ parameters) for the present task, PaRSP strategically reserves the vast majority of the parameter space. This prevents premature saturation and ensures sufficient capacity for the future, effectively harmonizing the stability-plasticity-sustainability triad.

### 5.3 Task Generalization

Beyond mitigating catastrophic forgetting on learned tasks, preserving the model's intrinsic reasoning capabilities on unseen tasks is vital. Following the evaluation protocol of Liao et al. (2025), we assess cross-task generalization across four representative benchmarks. While traditional CL paradigms often compromise general capabilities due to parameter drift, PaRSP achieves an effective balance between *continual adaptation* and *zero-shot generalization*.

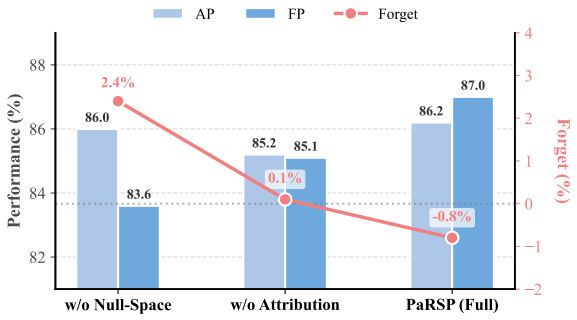


Figure 3: Ablation study (Llama-3.1-8B, SC Order 1). We compare **PaRSP** against variants lacking geometric constraints (**w/o Null-Space**) or using random masking (**w/o Attribution**).

As presented in Table 2, PaRSP consistently achieves superior performance across all metrics. *Notably*, while baseline methods exhibit significant performance degradation on reasoning-intensive tasks (specifically GSM8K and BBH), PaRSP successfully circumvents this regression. We attribute this success to our dual-constrained mechanism. By strictly confining parameter updates to task-specific subspaces, PaRSP preserves the synaptic stability of **generalist neurons** (as defined in Section 4.3). This confirms that our method acquires new task-specific skills without compromising the general intelligence of the backbone model.

### 5.4 Ablation Study

To rigorously quantify the individual contributions of each component within PaRSP, we conducted an ablation study on the Standard CL Benchmark (Order 1) using the Llama-3.1-8B backbone. We systematically removed or replaced key modules to verify the efficacy of *Task-Specific Parameter Region Activation* and *Orthogonal Update via Null Space Projection*. The comparative results are visualized in Figure 3.

**Effect of Task-Specific Parameter Region Activation.** To validate our neuron-level attribution, we substituted the task-specific mask with a random mask of equivalent sparsity (**w/o Attribution**). As shown in Figure 3, the decline in both FP and AP confirms that random selection fails to efficiently capture the "functional core." *Crucially*, while the drop in CL metrics appears moderate, further investigation reveals that random mask-

ing disproportionately impairs pre-trained **general knowledge** (detailed in Appendix E.3). This underscores that attribution-based localization is pivotal for both plasticity and stability: it targets the subspace most amenable to adaptation, whereas random allocation risks disrupting the *generalist neurons* underpinning the model's foundational capabilities.

**Effect of Orthogonal Update via Null Space Projection.** To assess the contribution of the geometric constraint, we removed the null-space projector while retaining the task-specific topological mask (denoted as **w/o Null-Space**). In this setting, active parameters are updated via standard gradient descent without orthogonality constraints. As shown in Figure 3, removing the projection has a negligible impact on plasticity (AP), indicating that the model can still learn current tasks effectively. However, it causes a catastrophic degradation in stability, evidenced by a sharp decline in FP. This outcome corroborates that while *spatial isolation* (masking) alone mitigates interference to a degree, it is insufficient for strict knowledge preservation. The **Null-Space Projection** serves as the decisive factor, providing the mathematical guarantee that new learning remains orthogonal to, and thus non-interfering with, established representations.

**Impact of Sparsity Ratio.** We further investigated the sensitivity of model performance to the active parameter ratio using Llama-3.1-8B on the Standard CL Benchmark (Order 1) (refer to Appendix E.2 for the detailed analysis). Empirical results identify a clear **functional saturation point** at approximately $\mathbf{5}\%$ sparsity. Beyond this threshold, performance gains exhibit *diminishing marginal returns*, confirming that dense updates are largely redundant. By calibrating the activation to this optimal subset, PaRSP strategically leaves $\sim \mathbf{95}\%$ of the parameter space untouched. This rigorous structural sparsity serves a dual purpose: it secures massive capacity for future sustainable evolution, and crucially, acts as a safeguard for the model's intrinsic capabilities. By confining updates to a minimal functional core, PaRSP avoids perturbing the vast majority of shared *generalist neurons*, thereby preserving the model's **General Intelligence** alongside task-specific adaptation.

### 5.5 Scalability Stress Test

To empirically validate the theoretical efficiency gains of the *Dual SVD* and *Implicit Projection* mechanisms detailed in Section 4.5, we conduct a computational stress test on the pre-computation overhead (SVD factorization and projector construction). We benchmark PaRSP across a spectrum of foundation models, scaling from 14B up to 70B parameters, on a single NVIDIA A100 (80GB) GPU. The calibration sample size is fixed at $N = 256$, and the sparsity ratio is set to $\rho \approx 17\%$ (corresponding to the LS benchmark setting).

| Model Backbone | $\mathbf{d_{hidden}}$ | $\mathbf{d_{inter}}$ | Total SVD Time (s) ↓ |
|---|---|---|---|
| Qwen-2.5-14B | 5,120 | 13,824 | 0.25 |
| Qwen-2.5-32B | 5,120 | 27,648 | 0.48 |
| **Llama-3.3-70B** | **8,192** | **28,672** | **0.71** |

Table 3: Wall-clock runtime for the entire PaRSP pre-computation phase across all active layers.

As shown in Table 3, runtime exhibits strictly linear scaling. For **Llama-3.3-70B**, our optimized pipeline completes the entire projection construction in merely **0.71s**. This one-off overhead constitutes less than **0.01%** of a typical training session. Furthermore, the *Implicit Gradient Projection* bypasses the instantiation of dense $\mathcal{O}(d^2)$ matrices, effectively preventing OOM errors on large-scale models. These results confirm that PaRSP is a production-ready, highly scalable method for industrial-scale LLMs.

## 6 Conclusion

In this work, we introduce the **PaRSP**, a bio-inspired paradigm designed to harmonize the stability-plasticity-sustainability triad in LLM continual learning. By synergizing *Task-Specific Parameter Region Activation* with *Orthogonal Update via Null Space Projection*, PaRSP enforces topological isolation to maximize plasticity for the **Present** and reserve capacity for the **Future**, while simultaneously imposing algebraic orthogonality to rigorously safeguard the **Past**. Crucially, to overcome the $\mathcal{O}(d^3)$ computational and $\mathcal{O}(d^2)$ memory bottlenecks inherent in massive LLMs, we introduce a **Dual SVD** optimization paired with an implicit projection mechanism. This mathematical restructuring shifts the asymptotic overhead to a linear scale, rendering the method highly tractable and production-ready for industrial-scale models. Extensive experiments validate that PaRSP achieves SOTA performance, offering a robust, scalable foundation for the sustainable evolution of general intelligence.

## Limitations

Despite the state-of-the-art performance achieved by PaRSP, we acknowledge several limitations that merit future investigation:

**Dependence on Training Task Boundaries.** While our evaluation setting is strictly task-agnostic during inference (no task IDs required), the training phase currently relies on explicit task boundaries to trigger the functional localization and null-space computation steps. This makes the method strictly applicable to *Task-Incremental* or *Domain-Incremental* settings. Extending PaRSP to a fully *Online Continual Learning* paradigm, where task boundaries are blurred or unknown during training, remains an open challenge.

**Potential Gradient Subspace Saturation.** Our method relies on the intersection of null spaces from historical tasks. Although the high dimensionality of LLMs provides a vast parameter space, theoretically, as the sequence of tasks grows indefinitely ($T \to \infty$), the available null space for updating parameters may eventually vanish (i.e., $\bigcap \mathcal{N}(\mathbf{\Sigma}_t) \to \emptyset$). This phenomenon, known as *subspace saturation*, could limit the model's plasticity for extremely long task sequences, necessitating mechanisms for dynamic capacity expansion or selective forgetting.

**Sensitivity to Calibration Quality.** The efficacy of our functional localization hinges on the representativeness of the calibration set used for neuron attribution. While we demonstrate robustness with a **sparse, fixed-budget sample size**, extreme distribution shifts between this small calibration set and the full task distribution could lead to suboptimal mask generation, potentially freezing neurons that are actually critical for the new task.

## Ethical Considerations

This work leverages exclusively publicly available datasets and models for technical advancements in continual learning. We foresee no direct ethical risks or privacy concerns beyond the inherent biases and limitations common to the foundational LLMs utilized.

## Acknowledgments

This work was supported by the Beijing Major Science and Technology Project (No. Z251100008125025), the National Natural Science Foundation of China (Nos. 62376270 and 62476060) and the Independent Research Project of the Key Laboratory of Cognition and Decision Intelligence for Complex Systems.

## References

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2019. Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations (ICLR)*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

OpenCompass Contributors. 2023. OpenCompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. 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, Dublin, Ireland. Association for Computational Linguistics.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. 2024. Loss of plasticity in deep continual learning. *Nature*, 632(8026):768–774.

Wenyu Du, Shuang Cheng, Tongxu Luo, Zihan Qiu, Zeyu Huang, Ka Chun Cheung, Reynold Cheng, and Jie Fu. 2024. Unlocking continual learning abilities in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6503–6522, Miami, Florida, USA. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daniela Gerz, Pei-Hao Su, Razvan Kusztos, Avishek Mondal, Michał Lis, Eshan Singhal, Nikola Mrkšić, Tsung-Hsien Wen, and Ivan Vulić. 2021. Multilingual and cross-lingual intent detection from spoken data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*,

pages 7468–7475, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5484–5495.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. 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 (CVPR)*.

Jinghan He, Haiyun Guo, Kuan Zhu, Zihan Zhao, Ming Tang, and Jinqiao Wang. 2024. 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 (EMNLP)*, pages 3254–3266, Miami, Florida, USA. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021,Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Gangwei Jiang, Caigao Jiang, Zhaoyi Li, Siqiao Xue, Jun Zhou, Linqi Song, Defu Lian, and Ying Wei. 2025. Unlocking the power of function vectors for characterizing and mitigating catastrophic forgetting in continual instruction tuning. In *The Thirteenth International Conference on Learning Representations*.

Sheena A Josselyn and Susumu Tonegawa. 2020. Memory engrams: Recalling the past and imagining the future. *Science*, 367(6473):eaaw4325.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Sangyoon Y. Ko, Yiming Rong, Adam I. Ramsaran, Xiaoyu Chen, Asim J. Rashid, Andrew J. Mocle, Jagroop Dhaliwal, Ankit Awasthi, Axel Guskjolen, Sheena A. Josselyn, and Paul W. Frankland. 2025. Systems consolidation reorganizes hippocampal engram circuitry. *Nature*, 643(8072):735–743.

Huanxuan Liao, Shizhu He, Yupu Hao, Jun Zhao, and Kang Liu. 2025. DATA: Decomposed attention-based task adaptation for rehearsal-free continual learning. *arXiv preprint arXiv:2502.11482*.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, pages 6467–6476. Curran Associates, Inc.

Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. 2022. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2252–2261.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In Gordon H. Bower, editor, *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual knowledge in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a Transformer. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Bruno A Olshausen and David J Field. 2004. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–487.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8024–8035. Curran Associates, Inc.

Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. 2023. Progressive prompts: Continual learning for language models. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Gobinda Saha, Isha Garg, and Kaushik Roy. 2021. Gradient projection memory for continual learning. In *International Conference on Learning Representations*.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. 2017. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, volume 30, pages 2990–2999. Curran Associates, Inc.

Olaf Sporns and Richard F. Betzel. 2016. Modular brain networks. *Annual Review of Psychology*, 67:613–640.

Fan-Keng Sun, Cheng-Hao Ho, and Hung-yi Lee. 2020. LAMOL: Language modeling for lifelong language learning. In *International Conference on Learning Representations (ICLR)*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Huajie Tan, Enshen Zhou, Zhiyu Li, Yijie Xu, Yuheng Ji, Xiansheng Chen, Cheng Chi, Pengwei Wang, Huizhu Jia, Yulong Ao, Mingyu Cao, Sixiang Chen, Zhe Li, Mengzhen Liu, Zixiao Wang, Shanyu Rong, Yaoxu Lyu, Zhongxia Zhao, Peterson Co, and 16 others. 2026. RoboBrain 2.5: Depth in sight, time in mind. *Preprint*, arXiv:2601.14352.

Yuqiao Tan, Shizhu He, Kang Liu, and Jun Zhao. 2025. Neural incompatibility: The unbridgeable gap of cross-scale parametric knowledge transfer in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21586–21601.

Rishabh Tiwari, Krishnateja Killamsetty, Rishabh Iyer, and Pradeep Shenoy. 2022. Gcr: Gradient coreset based replay buffer selection for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 99–108.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. 2023a. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10658–10671, Singapore. Association for Computational Linguistics.

Xiao Wang, Yuansen Zhang, Tianze Chen, Songyang Gao, Senjie Jin, Xianjun Yang, Zhiheng Xi, Rui Zheng, Yicheng Zou, Tao Gui, and 1 others. 2023b. TRACE: A comprehensive benchmark for continual learning in large language models. *arXiv preprint arXiv:2310.06762*.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11132–11152, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yequan Wang and Aixin Sun. 2025. Toward embodied agi: A review of embodied ai and the road ahead. *Preprint*, arXiv:2505.14235.

Yifan Wang, Yafei Liu, Chufan Shi, Haoling Li, Chen Chen, Haonan Lu, and Yujiu Yang. 2024. InsCL: A data-efficient continual learning paradigm for fine-tuning large language models with instructions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 663–677, Mexico City, Mexico. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tongtong Wu, Massimo Caccia, Zhuang Li, Yuan-Fang Li, Guilin Qi, and Gholamreza Haffari. 2022. Pretrained language model in continual learning: A comparative study. In *International Conference on Learning Representations*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Yiqun Yao, Xiang Li, Xin Jiang, Xuezhi Fang, Naitong Yu, Aixin Sun, and Yequan Wang. 2025. RoboEgo system card: An omnimodal model with native full duplexity. *Preprint*, arXiv:2506.01934.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *Computer Vision – ECCV 2016*, pages 69–85, Cham. Springer International Publishing.

Zeping Yu and Sophia Ananiadou. 2024. Neuron-level knowledge attribution in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3267–3280, Miami, Florida, USA. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, pages 649–657. Curran Associates, Inc.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. 2024. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9851–9873.

# A Complexity and Scalability Analysis

The rigorous deployment of a dual-constrained optimization method faces three primary computational hurdles: the potential overhead of the functional localization process, the cubic complexity of Singular Value Decomposition (SVD), and the linear growth of memory footprint for storing historical statistics. In this section, we demonstrate how our method systematically resolves these bottlenecks in a logical sequence, rendering the method computationally tractable.

## A.1 Minimal Overhead via Static Attribution

A potential concern regarding functional localization is the cost of identifying task-specific parameters. Unlike dynamic routing mechanisms (e.g., Mixture-of-Experts) that require gating computations for every token during both training and inference, our method employs a *Static Attribution Strategy*. The localization process is executed as a *one-shot pre-computation* step. We utilize a small, representative calibration set (typically a few hundred samples) to compute attribution scores. This data requirement is constant and does not scale linearly with the training set size, rendering the calibration overhead negligible—especially for large-scale datasets where this constitutes a minute fraction ($< 1\%$) of the data. This design effectively decouples *structure learning* (finding the mask) from *parameter learning* (updating weights). Consequently, the computational cost of attribution is amortized over thousands of training steps, rendering the per-step overhead negligible. Furthermore, once the binary mask $\mathbf{M}^{(k)}$ is generated, it remains frozen, allowing "non-active" regions to be computationally skipped, potentially accelerating training throughput.

## A.2 Computational Efficiency via Induced Sparsity

The primary computational bottleneck in null-space methods lies in the SVD operation, which typically scales with $\mathcal{O}(d^3)$ for dimension $d$. Naively applying this to all layers would be prohibitively expensive. However, we leverage the **structural sparsity** induced by the static attribution (Appendix A.1) to mitigate this cost. Since we only activate a sparse subset of parameters (the top-$k$ layers and specific neurons), the expensive SVD computations are strictly confined to these localized regions. The non-active layers require neither covariance collection nor projection calculation. Moreover, the SVD is performed solely during the initialization phase of each new task. It does not burden the iterative training loop, nor does it affect *inference latency*, as the projection constraints are implicitly baked into the updated weights.

## A.3 Memory Scalability via Recursive Aggregation

A critical challenge in continual learning is the "memory explosion" problem: storing independent covariance matrices for every previous task would lead to linear memory growth ($\mathcal{O}(T)$). To address this, we introduce a memory-constant optimization based on the algebraic properties of Positive Semi-Definite (PSD) matrices. We utilize the theorem that for any set of PSD matrices, the null space of their sum is exactly equal to the intersection of their individual null spaces:

$$\mathcal{N}\left(\sum_{t=1}^{T-1}\mathbf{\Sigma}_t\right) \equiv \bigcap_{t=1}^{T-1}\mathcal{N}(\mathbf{\Sigma}_t) \tag{8}$$

Based on this, we implemented a *Stream Accumulation Strategy*. Instead of loading multiple matrix files, we recursively add the uncentered covariance matrix of each historical task to a single accumulator buffer. This reduces the memory complexity from $\mathcal{O}(T \cdot d^2)$ to a constant $\mathcal{O}(d^2)$. Regardless of the number of tasks, we only need to maintain a single aggregated uncentered covariance matrix, ensuring infinite scalability.

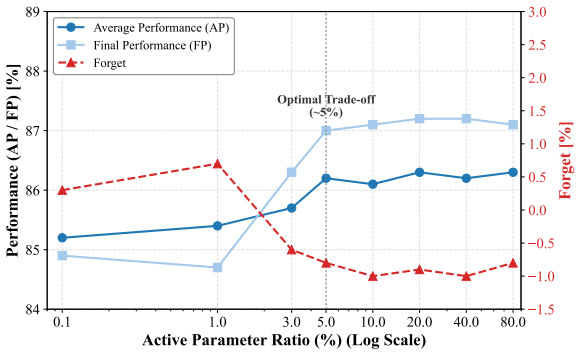


Figure 4: Sensitivity analysis of the active parameter ratio on the Standard CL Benchmark (Order 1) using **Llama-3.1-8B**. The plot reveals a clear **functional saturation point** at $\sim$ **5**% sparsity (marked by the vertical line). Beyond this threshold, performance gains (AP/FP) exhibit diminishing marginal returns, while the Forgetting Rate stabilizes. This empirically validates that activating only a sparse functional core is sufficient for plasticity, effectively reserving $\sim 95\%$ of the parameter space for future sustainability.

## B Details of Hierarchical Selection Process

As discussed in Section 4.2, we execute functional localization in two stages to optimize the trade-off between computational overhead and identification accuracy:

- **Stage 1: Layer-wise Filtering (Coarse-grained).** We first aggregate the attribution scores (defined in Equation (3)) for the Multi-Head Self-Attention (MHSA) and Feed-Forward Network (FFN) blocks across all layers. The layer-level importance is defined as the sum of attribution scores of all constituent neurons or heads. Layers that exhibit negligible or negative contributions are deemed task-irrelevant and excluded from the active scope. We retain the top-$L_s$ layers based on these cumulative attribution scores.
- **Stage 2: Neuron-wise Identification (Fine-grained).** Within the selected top-$L_s$ layers, we further scrutinize individual components. For FFNs, we compute scores for each intermediate neuron; for MHSA, we evaluate each attention head. By iterating through these dimensions, we isolate the specific neurons that maximize the attribution score.

The resulting indices form the binary mask $\mathbf{M}^{(k)}$, where an entry $\mathbf{M}_i^{(k)} = 1$ indicates an active parameter and $\mathbf{M}_i^{(k)} = 0$ indicates a frozen one.

## C Gradient Rectification Implementation

As discussed in Section 4.4, implementing the dual constraints directly during the forward pass would be computationally expensive. Instead, we implement an efficient *Gradient Rectification* strategy using PyTorch backward hooks.

**Hook Registration Mechanism.** For each identified task-relevant layer, we register two sequential backward hooks that modify the gradient flow before the optimizer step:

- **Mask Hook (Topological Gate):** This hook filters the gradient flow based on the neuron indices derived from the attribution stage. It performs an element-wise multiplication with the binary mask $\mathbf{M}_l$, effectively zeroing out gradients for frozen neurons:

$$\frac{\partial \mathcal{L}}{\partial w_{ij}} \leftarrow 0 \quad \text{if} \quad (\mathbf{M}_l)_{ij} = 0$$

- **Projector Hook (Geometric Filter):** This hook performs a right-multiplication of the gradient matrix with the pre-computed null-space projector $\mathbf{P}_l$. This operation rectifies the gradient direction to ensure orthogonality with the old task subspace:

$$\Delta \mathbf{W} \perp \text{span}(\mathbf{X}_{old})$$

**Computational Efficiency.** By decoupling constraint computation (a one-time cost per task) from the iterative training loop, and applying corrections only during the backward pass, our method introduces negligible overhead to training throughput while providing theoretical guarantees against catastrophic forgetting.

## D Detailed Experimental Setup

This appendix provides comprehensive specifications for the datasets, task configurations, and implementation protocols to ensure full reproducibility of our results.

### D.1 Datasets and Benchmarks

**Training Tasks.** We provide comprehensive specifications for the datasets utilized in our continual learning experiments. The **Standard CL Benchmark (SC)** (Zhang et al., 2015) consists of four text classification datasets: AG News, Amazon Reviews, DBpedia, and Yahoo Answers, as selected by Wang et al. (2023a). The **Long Sequence**

| Dataset Name | Category | Task | Domain | Metric |
|---|---|---|---|---|
| Yelp | CL Benchmark | Sentiment Analysis | Yelp Reviews | Accuracy |
| Amazon | CL Benchmark | Sentiment Analysis | Amazon Reviews | Accuracy |
| DBpedia | CL Benchmark | Topic Classification | Wikipedia | Accuracy |
| Yahoo | CL Benchmark | Topic Classification | Yahoo Q&A | Accuracy |
| AG News | CL Benchmark | Topic Classification | News | Accuracy |
| MNLI | GLUE | Natural Language Inference | Various | Accuracy |
| QQP | GLUE | Paragraph Detection | Quora | Accuracy |
| RTE | GLUE | Natural Language Inference | News, Wikipedia | Accuracy |
| SST-2 | GLUE | Sentiment Analysis | Movie Reviews | Accuracy |
| WiC | SuperGLUE | Word Sense Disambiguation | Lexical Databases | Accuracy |
| CB | SuperGLUE | Natural Language Inference | Various | Accuracy |
| COPA | SuperGLUE | Question and Answering | Blogs, Encyclopedia | Accuracy |
| BoolQA | SuperGLUE | Boolean Question and Answering | Wikipedia | Accuracy |
| MultiRC | SuperGLUE | Question and Answering | Various | Accuracy |
| IMDB | SuperGLUE | Sentiment Analysis | Movie Reviews | Accuracy |

Table 4: Detailed specifications of the 15 classification datasets comprising the Long Sequence (LS) benchmark (Razdaibiedina et al., 2023). The initial five tasks originate from the standard CL benchmark protocol (Zhang et al., 2015).

**Benchmark (LS)** (Razdaibiedina et al., 2023) extends this scope to comprise 15 diverse tasks, incorporating GLUE and SuperGLUE datasets. Detailed statistics and metrics for all 15 datasets (encompassing both SC and LS) are enumerated in Table 4. Regarding data partitioning, for both SC and LS benchmarks, we randomly sample **1,000 instances** per task for training and reserve **500 instances** per task for validation and testing. The **TRACE Benchmark** (Wang et al., 2023b) includes 8 challenging tasks designed to evaluate LLM-specific capabilities such as code generation and reasoning (details in Table 10). To ensure adequate adaptation for these complex tasks, we utilize **5,000 training instances** per task.

**Generalization Benchmarks.** To evaluate the model's cross-task generalization capabilities on unseen domains, we employ four widely recognized benchmarks:

- **MMLU** (Hendrycks et al., 2021): The Massive Multitask Language Understanding benchmark covers 57 subjects across STEM, the humanities, and social sciences, ranging from elementary to professional levels. It serves as a comprehensive test of world knowledge and problem-solving ability.
- **GSM8K** (Cobbe et al., 2021): A dataset of 8.5K high-quality, linguistically diverse grade school math word problems. It requires the model to perform multi-step elementary mathematical reasoning to derive the correct answer.
- **BBH** (Suzgun et al., 2023): BIG-Bench Hard (BBH) is a subset of 23 challenging tasks from the BIG-Bench suite, spanning diverse areas such as arithmetic and symbolic reasoning.
- **AGIEval** (Zhong et al., 2024): A human-centric benchmark derived from high-standard official admission and qualification exams. It assesses the model's ability to tackle tasks designed for human intelligence.

### D.2 Task Sequence Configurations

The specific task execution sequences utilized in our continual learning experiments are enumerated in Table 11.

### D.3 Implementation Details

We provide a comprehensive overview of the software environment, hyperparameters, and configuration protocols.

**Software Framework.** Our code implementation leverages the Hugging Face `transformers` library (v4.46.2) (Wolf et al., 2020) and the PyTorch framework (v2.9.1+cu128) (Paszke et al., 2019). For the evaluation of cross-task generalization on unseen datasets, we utilize the OpenCompass toolkit (Contributors, 2023), adhering strictly to its default configuration protocols.

**Training and Hyperparameters.** All models were trained using the AdamW optimizer. To ensure stable convergence, we employed a linear learning rate scheduler with warmup and linear decay. A consistent *early stopping* mechanism was applied across all tasks to prevent overfitting, with the maximum training duration capped at specific epochs depending on the benchmark complexity.
**Benchmark-Specific Configurations.** We tailored the hyperparameters and the PaRSP activation ratio (sparsity) according to the difficulty and scale of each benchmark:

- **Standard CL (SC) & Long Sequence (LS):** Since SC represents simpler classification tasks, we set a stricter sparsity constraint ($\sim \mathbf{5}\%$ active parameters) to maximize stability. For the more diverse LS benchmark, we increased the activation ratio to $\sim \mathbf{17}\%$ to accommodate broader knowledge acquisition. The maximum training duration was set to 3 epochs for both benchmarks.
- **TRACE:** Given the challenging nature of reasoning and coding tasks in TRACE, we maintained the $\sim \mathbf{17}\%$ activation ratio but increased the maximum training duration to 5 epochs to ensure sufficient adaptation.

Specific learning rates (LR) for each model architecture (Llama-3.1-8B and Qwen-2.5-7B) varied across benchmarks to ensure optimal performance. The detailed hyperparameter settings are enumerated in Table 5.

| Benchmark | Model | LR | Epochs | Sparsity |
|---|---|---|---|---|
| Standard CL (SC) | Llama-3.1-8B | 3e-6 | 3 | $\sim 5\%$ |
| | Qwen-2.5-7B | 3e-5 | 3 | $\sim 5\%$ |
| Long Sequence (LS) | Llama-3.1-8B | 3e-6 | 3 | $\sim 17\%$ |
| | Qwen-2.5-7B | 3e-5 | 3 | $\sim 17\%$ |
| TRACE | Llama-3.1-8B | 1e-5 | 5 | $\sim 17\%$ |
| | Qwen-2.5-7B | 2e-5 | 5 | $\sim 17\%$ |

Table 5: Hyperparameter configurations for different benchmarks and backbone models. LR: Learning Rate; Epochs: Maximum training epochs per task.

**PaRSP-Specific Protocols.** The implementation of PaRSP involves two critical data sampling phases. We adopted a unified and efficient sampling strategy to minimize computational overhead:

- **Task-Specific Activation (Calibration):** During the functional localization phase for a new task $T_k$, we randomly sampled **256 instances** from the training set of $T_k$. These samples serve as the calibration data to calculate neuron attribution scores (Equation (3)), determining the binary mask $\mathbf{M}^{(k)}$.
- **Null-Space Projection (Covariance):** To construct the projector $\mathbf{P}$, we maintain historical statistics without storing the full dataset. For each previous task $T_j (j < k)$, we randomly sampled **256 instances** to compute its uncentered covariance matrix $\mathbf{\Sigma}_j$. These matrices are recursively aggregated to derive the orthogonal null space, ensuring strict protection of past knowledge.

**Preservation of General Intelligence (Task 0 Initialization).** Beyond mitigating forgetting for learned tasks, ensuring the stability of the model's pre-trained general capabilities is paramount. We treat the preservation of broad world knowledge and reasoning skills as the *zeroth task* ($T_0$). Prior to the commencement of the continual learning sequence, we construct a compact **General Knowledge Anchor Set** by sampling **256 instances** from the MMLU benchmark (Hendrycks et al., 2021). To ensure representativeness within a limited budget, we select seven diverse subtasks covering two cognitive dimensions:

- *Knowledge-Intensive Domains* (Global Facts, High School World History, Psychology, Sociology) to anchor static declarative memory.
- *Reasoning-Intensive Domains* (Elementary Mathematics, Biology, Conceptual Physics) to preserve the model's cognitive upper bound for multi-step reasoning.

Using this anchor set, we compute an initial uncentered covariance matrix $\mathbf{\Sigma}_{gen}$. During the subsequent training of all CL tasks, the null space of $\mathbf{\Sigma}_{gen}$ is incorporated into the orthogonal projection constraint. This effectively locks the parameter subspace responsible for general intelligence, ensuring that subsequent model updates $\Delta\mathbf{W}$ remain orthogonal to, and thus non-destructive towards, the model's intrinsic capabilities.

# E Supplementary Experimental Results

## E.1 Task-wise Performance across All Permutations

We present a comprehensive breakdown of the experimental results for each individual task order

| Methods | Standard CL Benchmark (SC) | | | | Long Sequence Benchmark (LS) | | | | TRACE |
|---|---|---|---|---|---|---|---|---|---|
| | Order 1 | Order 2 | Order 3 | Avg | Order 4 | Order 5 | Order 6 | Avg | Order 7 |
| *# Llama-3.1-8B* | | | | | | | | | |
| SeqLoRA | 79.9 | 79.0 | 80.0 | 79.6 | 74.2 | 73.7 | 76.5 | 74.8 | 65.1 |
| LoRAReplay | 80.1 | 80.6 | 80.1 | 80.3 | 83.2 | 80.7 | 82.2 | 82.0 | **78.7** |
| O-LoRA (Wang et al., 2023a) | 71.8 | 72.2 | 72.8 | 72.3 | 73.1 | 69.4 | 71.6 | 71.4 | 36.7 |
| DATA (Liao et al., 2025) | 81.4 | 80.7 | 80.5 | 80.9 | 80.7 | 77.7 | 81.5 | 80.0 | 72.7 |
| + Replay | 80.6 | 81.0 | 80.7 | 80.8 | 83.1 | 81.7 | 81.8 | 82.2 | 77.6 |
| **PaRSP (Ours)** | **87.0** | **86.7** | **85.2** | **86.3** | **89.6** | **87.3** | **87.2** | **88.0** | 73.8 |
| *# Qwen-2.5-7B* | | | | | | | | | |
| SeqLoRA | 80.0 | 77.9 | 78.4 | 78.8 | 79.5 | 79.1 | 81.1 | 79.9 | 65.1 |
| LoRAReplay | 80.7 | 80.6 | 80.1 | 80.5 | 83.3 | 83.2 | 82.7 | 83.1 | 75.7 |
| DATA (Liao et al., 2025) | 79.8 | 79.1 | 79.4 | 79.4 | 79.8 | 80.2 | 81.5 | 80.5 | 70.4 |
| + Replay | 80.3 | 80.6 | 79.9 | 80.3 | 83.7 | 82.9 | 82.9 | 83.2 | **77.3** |
| **PaRSP (Ours)** | **86.0** | **87.1** | **86.4** | **86.5** | **89.3** | **87.5** | **88.1** | **88.3** | 73.2 |

Table 6: Summary of Final Performance (FP) on Standard CL, Long Sequence, and TRACE benchmarks using Llama-3.1-8B and Qwen-2.5-7B backbones. We report the averaged accuracy after training on the final task. **Bold** denotes the best performance.

across the three benchmarks in Table 6. Consistently across all task permutations, our proposed **PaRSP** demonstrates superior robustness, effectively mitigating catastrophic forgetting (CF) while preserving high plasticity for new task adaptation.

### E.2 Sensitivity Analysis of Active Parameter Ratio

To empirically determine the optimal equilibrium between plasticity (learning the present) and sustainability (reserving for the future), we conducted a fine-grained sensitivity analysis on the activation ratio of the PaRSP spatial mask. We employed the Llama-3.1-8B backbone on the Standard CL Benchmark (Order 1) and varied sparsity levels by adjusting selection thresholds for both layers and neurons. The results are visualized in Figure 4.

**Regime 1: Robustness at Extreme Sparsity.** As illustrated in Figure 4, even in the regime of extreme sparsity—activating merely $\sim \mathbf{0.1}\%$ of parameters—the model exhibits remarkable plasticity rather than the expected collapse. The Average Performance (AP) and Final Performance (FP) remain robust at **85.2%** and **84.9%**, respectively. We attribute this surprising efficacy to the precision of our attribution mechanism. It suggests that LLMs inherently possess the latent knowledge required for these tasks; the attribution mask successfully isolates the critical "functional neurons" needed to activate these capabilities with minimal parameter perturbation. This implies that for pre-trained LLMs, massive parameter updates are often unnecessary for task adaptation.

**Regime 2: Performance Saturation.** As we relax the sparsity constraint, performance improves marginally. When the activation ratio reaches approximately **5%**, AP and FP rise to **86.2%** and **87.0%**, respectively. Crucially, as the activation ratio exceeds this threshold, we observe a clear phenomenon of *diminishing marginal returns*: further increases in parameter usage yield negligible performance gains. This plateau indicates a "functional saturation point," suggesting that mobilizing just **5%** of the network's capacity is sufficient to achieve near-optimal adaptation for tasks of this complexity.

**Implications for Efficiency and Sustainability.** These findings provide compelling empirical evidence exposing the redundancy of current dense training paradigms. Updating 100% of parameters for every new task is computationally inefficient and detrimental to stability, as it disturbs parameters that contribute little to the current task but may be vital for past knowledge. By identifying this saturation point, PaRSP effectively circumvents these pitfalls. It activates the necessary $\sim \mathbf{5}\%$ functional core to maximize plasticity for the present, while strategically preserving the remaining $\sim \mathbf{95}\%$ of the network as an immutable resource. This rigorous preservation of parameter space is the core mechanism that secures the model's future sustainability for long-term evolution.

| Method | Mask Strategy | MMLU (Avg. Acc) |
|---|---|---|
| **PaRSP (Ours)** | **Attribution-based** | **0.876** |
| w/o Attribution | Random Selection | 0.830 |
| | *Relative Degradation* | *-5.25%* |

Table 7: Impact of masking strategy on general knowledge retention. Evaluated on Llama-3.1-8B after SC Order 1 training (500 samples total). The random mask causes implicit damage to general capabilities compared to PaRSP's targeted updates.

### E.3 Impact of Task-Specific Activation on General Knowledge Preservation

In our ablation study (Section 5.4), we observed a counter-intuitive phenomenon: replacing the *Task-Specific Parameter Region Activation* module with a random mask (denoted as **w/o Attribution**) resulted in only a moderate decline in CL metrics (FP and AP) on the Standard CL Benchmark (Order 1). This observation necessitates a deeper investigation into the *hidden costs* of random parameter allocation.

**Theoretical Analysis: Blind vs. Targeted Plasticity.** To understand this discrepancy, we revisit the core philosophy of PaRSP. The essence of our localization strategy is to pinpoint **Specialized Neurons**—those most sensitive to the current task—and grant them plasticity. This targeted approach serves a dual purpose: it maximizes performance for the present task while freezing **Generalist Neurons** to preserve the model's broad cognitive capabilities (the past). In contrast, a random mask, while maintaining the same sparsity ratio, lacks this semantic awareness. It indiscriminately updates parameters. Although the model can still force these randomly selected neurons to overfit the current task, this "blind" update trajectory risks overwriting critical, task-agnostic knowledge structures encoded during pre-training. We hypothesize that this leads to an *implicit degradation* of the model's general intelligence.

**Experimental Setup: Quantifying Implicit Damage.** To validate this hypothesis, we evaluated the models on the Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2021). We carefully selected seven representative subtasks to form a robust evaluation set specifically tailored to the capabilities of the Llama-3.1-8B model. These subjects span the spectrum from knowledge retrieval to complex reasoning:

- **Knowledge-Intensive (Comfort Zone):** *Global Facts*, *High School World History*, *High School Psychology*, and *Sociology*. These subjects primarily test the retention of static world knowledge.
- **Reasoning-Intensive (Capability Boundary):** *Elementary Mathematics*, *High School Biology*, and *Conceptual Physics*. These subjects require multi-step reasoning and reflect the model's cognitive upper bound.

We conducted the evaluation on the Llama-3.1-8B model after it had completed the training of Standard CL (Order 1). To ensure efficient yet representative evaluation, we constructed a test set containing a total of 500 samples, randomly aggregated from the seven selected subjects.

**Results and Discussion.** The comparative results are summarized in Table 7. The data reveals a distinct performance gap masked by standard CL metrics. The model utilizing the random mask suffered a significant regression in general capabilities, with the average accuracy across the test set dropping from **0.876** (Standard PaRSP) to **0.830**. This empirical evidence confirms that while random sparse updates can technically accommodate new tasks, they do so at the expense of the model's foundational knowledge. By removing the task-specific activation module, the updates impinge upon generalist neurons, eroding the model's pre-trained capabilities. In conclusion, the *Task-Specific Parameter Region Activation* component is indispensable not only for optimizing plasticity but, more critically, for ensuring the safety and stability of the model's general intelligence during continual evolution.

## F Extension to Multimodal Embodied AI: Preserving Foundational Perception

To demonstrate the architecture-agnostic universality of our method, we extend the application of PaRSP to the domain of Multimodal Large Language Models (MLLMs) for Embodied AI.

**Motivation and The "Cognitive Anchor" Protocol.** Embodied agents deployed in physical environments face a continuous stream of multimodal tasks. An agent must sequentially acquire advanced capabilities, such as **visual grounding** and **affordance grounding**, without catastrophically forgetting its foundational visual-linguistic perception. Traditional fine-tuning in such cross-modal settings

often exacerbates parameter drift, leading to severe capability degradation. To investigate this interference, we utilize the **RoboBrain-2.5-4B** (Tan et al., 2026) foundation model within a targeted **Cognitive Anchor** protocol. Specifically, we designate General VQA (Task 1) as a zero-shot anchor to quantify the degradation of foundational visual intelligence following the sequential acquisition of **Visual Grounding** (Task 2) and **Affordance Grounding** (Task 3). This untrained anchor serves as a persistent probe to measure model stability amidst a stream of specialized embodied tasks.

**Curriculum and Method Adaptation.** The continuous learning stream consists of:

- **Task 1 (Anchor):** General VQA evaluated on the VQAv2 benchmark (Goyal et al., 2017).
- **Task 2 (Learning):** Visual Grounding utilizing the RefCOCO and RefCOCO+ datasets (Yu et al., 2016).
- **Task 3 (Learning):** Affordance Grounding across diverse scenes utilizing the **AGD20K** dataset (Luo et al., 2022).

We compare standard Full Fine-Tuning against **PaRSP**, utilizing **500** randomly sampled instances for Task 1 and **1,000** per task for Task 2 and Task 3. Given the architectural complexity of interleaving visual encoders and LLM backbones, we focus entirely on the *Task-Specific Parameter Region Activation* component. Prior to learning Tasks 2 and 3, we utilize 256 multimodal calibration samples to compute attribution scores. We strictly enforce topological isolation by activating only a sparse $\sim \mathbf{5}\%$ functional core for spatial adaptation, preserving the remaining $95\%$ of the network as an immutable "visual cognition memory bank."

**Results and Mechanistic Insights.** The impact of acquiring embodied skills on the foundational VQA capability is quantified in Table 8.

| Method | VQA Accuracy | Relative Drop (↓) |
|---|---|---|
| **Zero-Shot** | 74.6% | - |
| Full Fine-Tuning | 68.4% | -6.2% |
| **PaRSP (Ours)** | **72.2%** | **-2.4%** |

Table 8: Degradation of foundational visual perception (VQAv2) after sequentially learning Visual Grounding and Affordance Grounding on RoboBrain-2.5-4B.

As evident in Table 8, standard Full Fine-Tuning causes significant parameter drift, plunging the foundational VQA accuracy from 74.6% to 68.4%. This confirms that unconstrained adaptation to spatial coordinates overwrites the generalist neurons responsible for high-level visual reasoning. Conversely, PaRSP effectively circumvents this semantic degradation, maintaining a VQA accuracy of 72.2%. This case study empirically validates our core bio-inspired philosophy: by topologically disentangling "action/spatial neurons" from "cognitive neurons," PaRSP enables the sustainable evolution of multimodal agents without sacrificing their foundational intelligence.

## G Mechanistic Probing of Functional Partitioning

A core theoretical premise of PaRSP is that foundation models intrinsically compartmentalize knowledge into sparse, functionally disentangled subnetworks. To empirically validate this hypothesis, we conduct a mechanistic probing study on **FLM-Audio** (Yao et al., 2025; Wang and Sun, 2025), a speech-language large model that processes continuous audio and text streams within a unified parameter space. This dense architectural fusion inherently exacerbates the risk of cross-modal interference, making it an exceptionally rigorous testbed. Our primary objective is to uncover the innate *representational disentanglement* between acoustic perception and semantic reasoning. Demonstrating that distinct modalities naturally occupy disjoint parameter subspaces provides the definitive physical justification for PaRSP's core mechanism: topologically isolating task-specific circuits while mathematically protecting shared cognitive bridges.

**Probing Setup and "Cognitive Anchor" Protocol.** We aim to determine whether the neurons responsible for *Acoustic Perception* and *Semantic Reasoning* are physically entangled or topologically isolated. We establish two distinct task anchors:

- **Semantic Anchor (Text):** We randomly sample 256 textual instances from the **GSM8K** benchmark (Cobbe et al., 2021) to activate high-level logical reasoning circuits.
- **Acoustic Target (Audio):** We randomly sample 256 speech instances from the **MINDS-14** dataset (Gerz et al., 2021) to activate auditory processing circuits.

Using our static attribution method (Equation (3)), we compute the attribution scores for both the Feed-

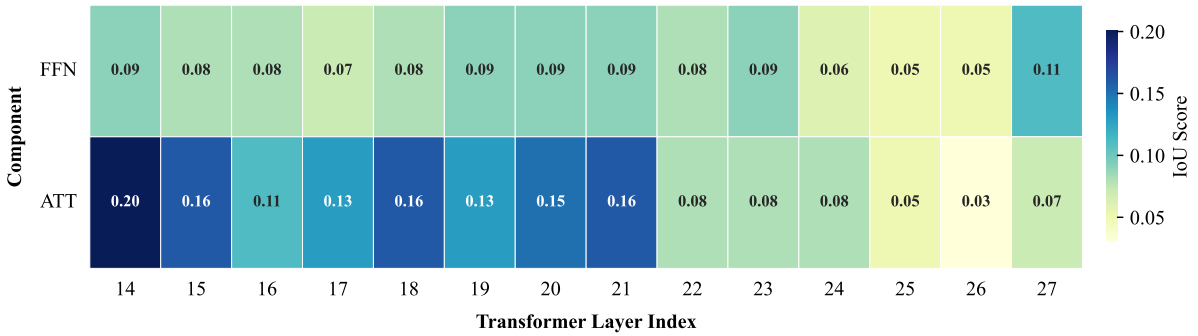


Figure 5: Layer-wise Modality Disentanglement IoU (Text vs. Audio) for FFN and ATT components. The downward trend illustrates the transition from shared routing to specialized storage.

Forward Network (FFN) modules and Attention (ATT) modules across the mid-to-deep layers (Layers 14–27), which are critical for knowledge storage and routing. We isolate the top-5% most salient neurons for each modality and quantify their spatial overlap using the Intersection over Union (IoU) metric.

**Global Disentanglement Analysis.** The global intersection statistics are summarized in Table 9. The remarkably low IoU scores—**8.00%** for FFN modules and **11.18%** for ATT modules—demonstrate an extreme degree of modality disentanglement. Crucially, the observation that FFN modules exhibit a lower overlap than ATT modules aligns with prevailing mechanistic interpretability theories: ATT modules primarily act as *information routers* aggregating context, necessitating a baseline level of shared processing pathways across modalities; conversely, FFN modules function as localized *key-value memories* (Geva et al., 2021), requiring extreme specialization to avoid semantic collision between acoustic features and logical rules. Consequently, unconstrained full-parameter fine-tuning on speech tasks inevitably overwrites these disjoint semantic memory banks, validating the necessity of PaRSP's topological isolation to prevent catastrophic reasoning collapse.

| Component | Modality Disentanglement IoU (↓) |
|---|---|
| **FFN (Storage)** | **8.00%** |
| **ATT (Routing)** | **11.18%** |

Table 9: Global modality disentanglement statistics (Layers 14–27, Top-5% activation) on FLM-Audio. A lower IoU indicates higher physical isolation between semantic and acoustic circuits.

**Layer-wise Analysis: The "Cognitive Funnel" Phenomenon.** To dissect the internal representation flow, we visualize the layer-wise IoU in Figure 5. We observe a distinct and monotonic *"Cognitive Funnel"* trend:

- **Shallow Layers (Shared Routing):** The ATT modules in earlier layers (e.g., Layer 14) exhibit relatively higher structural overlap (IoU = 0.20). This indicates their role as multimodal feature aggregators, where acoustic and textual streams share foundational processing hubs before divergence.
- **Deep Layers (Specialized Storage):** As depth progresses, the parameters become profoundly specialized. By Layer 26, the ATT IoU plunges to **0.03**, and the FFN IoU to **0.05**, effectively functioning as strictly disjoint, modality-specific memory banks.

**Mechanistic Implications for PaRSP.** These probing results provide profound physical grounding for PaRSP's dual-constrained design. Because deep layers are highly independent, *Spatial Masking* alone is highly effective at isolating new task learning. However, because shallow layers harbor "Generalist Neurons" (the shared multimodal hubs), relying solely on structural sparsity is dangerously insufficient. Therefore, the *Null-Space Orthogonal Update* becomes mathematically imperative to shield these shared shallow circuits from gradient interference. This empirical evidence confirms that PaRSP is not an artificial algorithmic constraint, but a natural alignment with the intrinsic circuit organization of foundation models.

| Dataset | Source | Avg len | Metric | Language | #Data |
|---|---|---|---|---|---|
| *Domain-specific* | | | | | |
| ScienceQA | Science | 210 | Accuracy | English | 5,000 |
| FOMC | Finance | 51 | Accuracy | English | 5,000 |
| MeetingBank | Meeting | 2853 | ROUGE-L | English | 5,000 |
| *Multi-lingual* | | | | | |
| C-STANCE | Social media | 127 | Accuracy | Chinese | 5,000 |
| 20Minuten | News | 382 | SARI | German | 5,000 |
| *Code Completion* | | | | | |
| Py150 | Github | 422 | Edim Similarity | Python | 5,000 |
| *Mathematical Reasoning* | | | | | |
| NumGLUE-cm | Math | 32 | Accuracy | English | 5,000 |
| NumGLUE-ds | Math | 21 | Accuracy | English | 5,000 |

Table 10: Statistical summary of the TRACE benchmark (Wang et al., 2023b). "Source" signifies context origin, while "Avg len" denotes mean length in words (for English, German, and code) or characters (for Chinese). SARI is a specialized metric employed for simplification tasks.

| Benchmark | Order | Task Sequence |
|---|---|---|
| Standard CL Benchmark | 1 | dbpedia → amazon → yahoo → ag |
| | 2 | dbpedia → amazon → ag → yahoo |
| | 3 | yahoo → amazon → ag → dbpedia |
| Long Sequence Benchmark | 4 | mnli → cb → wic → copa → qqp → boolqa → rte → imdb → yelp → amazon → sst-2 → dbpedia → ag → multirc → yahoo |
| | 5 | multirc → boolqa → wic → mnli → cb → copa → qqp → rte → imdb → sst-2 → dbpedia → ag → yelp → amazon → yahoo |
| | 6 | yelp → amazon → mnli → cb → copa → qqp → rte → imdb → sst-2 → dbpedia → ag → yahoo → multirc → boolqa → wic |
| TRACE | 7 | c-stance → fomc → meetingbank → py150 → scienceqa → numglue-cm → numglue-ds → 20minuten |

Table 11: The experiments utilize seven task orderings categorized into three groups: Orders 1–3 follow the Standard CL Benchmark (Zhang et al., 2015); Orders 4–6 involve the Long Sequence Benchmark spanning 15 diverse tasks (Razdaibiedina et al., 2023); and Order 7 represents the TRACE benchmark, which comprises eight datasets (Wang et al., 2023b).