ARIA: Adaptive Recurrent Intelligence Architecture

Morphogenic Attention and Slow-Pathway Consolidation for Continual Learning

Darshan Poudel
Independent Researcher
poudeldarshan44@gmail.com

June 2026 (Preprint)

Abstract

Catastrophic forgetting remains a fundamental obstacle in sequential learning. Most existing mitigations, including Elastic Weight Consolidation (EWC), replay buffers, and progressive networks, protect weights without changing how the model allocates capacity across tasks. We introduce ARIA, an architecture in which the capacity allocation *is* the learning signal. ARIA combines four jointly-trained mechanisms: (1) **Morphogenic Attention (MA)**, where heads split or merge based on learned viability scores; (2) **Plasticity-Gated MLP (PG-MLP)**, a dual fast/slow pathway whose gate routes each token toward volatile or consolidated representations; (3) **Architecture Genome Vector (AGV)**, a global latent vector that conditions all blocks via FiLM affine transforms; and (4) **Cognitive Budget Allocator (CBA)**, which predicts per-layer compute budgets from input complexity statistics. We additionally introduce **Slow-Pathway Consolidation (SPC)**, an EWC-style Fisher regulariser applied *only* to slow-pathway weights, allowing fast pathways to adapt freely while protecting consolidated knowledge. We further introduce three lightweight enhancements, Slow-Pathway Activation Distillation (SPAD), task-conditioned gate routing, and asymmetric learning rates, that together constitute ARIA-v2. On Split-MNIST (5 seeds, parameter-matched baselines), ARIA+SPC achieves **98.50%** average accuracy with **0.86%** forgetting, outperforming EWC (97.35%, 2.29%) by +1.15 points accuracy and reducing forgetting by **62%** relative, while requiring 50% fewer Fisher parameters than EWC.

1 Introduction

A learning system that encounters tasks sequentially faces a dilemma: plasticity (rapid adaptation to new tasks) and stability (retention of old tasks) are in direct tension [McCloskey and Cohen, 1989, French, 1999]. Neural networks trained by stochastic gradient descent exhibit extreme plasticity at the cost of stability, a phenomenon known as *catastrophic forgetting*.

The continual learning literature has addressed this through three broad strategies: regularisation (EWC [Kirkpatrick et al., 2017], SI [Zenke et al., 2017]), replay (DER++ [Buzzega et al., 2020], ER [Rolnick et al., 2019]), and progressive architectures (PNN [Rusu et al., 2016], PackNet [Mallya and Lazebnik, 2018]). Each strategy has a known failure mode: regularisation is too conservative for long task sequences, replay has memory cost, and progressive networks do not transfer across tasks.

ARIA takes a fourth approach: the architecture adapts its *structure* in response to task demands, distributing capacity dynamically so that both new and old knowledge can coexist without a fixed trade-off. The core insight is that the fast/slow dichotomy observed in biological memory [McClelland et al., 1995] can be implemented as a learned gate per neuron, with slow weights receiving reduced gradient updates during high-plasticity phases and stronger Fisher-based protection after consolidation.

Contributions.

1. **Morphogenic Attention**: the first attention mechanism where head count is a learned, dynamic quantity, with principled split (specialisation) and merge (consolidation) operations governed by viability scores and cosine similarity.
2. **Plasticity-Gated MLP**: a dual-pathway MLP with a per-token learned gate and bimodal specialisation loss, inspired by complementary learning systems.
3. **Slow-Pathway Consolidation (SPC)**: targeted Fisher regularisation over slow-pathway weights only, using 50% fewer Fisher parameters than EWC while improving specificity.

4. **Architecture Genome Vector**: a shared latent vector that decodes to skip probabilities, attention temperature, and FiLM scale/shift, conditioning all blocks with a single differentiable structural descriptor.

5. **Cognitive Budget Allocator**: a lightweight network predicting per-layer compute allocation from raw input statistics.

2 Related Work

Regularisation methods. EWC [Kirkpatrick et al., 2017] penalises movement of weights that were important for past tasks, measured by the diagonal Fisher information matrix. SI [Zenke et al., 2017] accumulates importance online during training. Both apply uniform regularisation across all parameters, including weights designed to be volatile, which conflicts with plasticity-gated architectures like ARIA.

Replay methods. DER++ [Buzzega et al., 2020] stores (input, logit) pairs and penalises divergence from stored predictions. ER [Rolnick et al., 2019] stores raw examples. Replay methods require a memory buffer whose size scales with the number of tasks. In our experiments, DER++ with a small buffer (200 samples) underperforms EWC significantly.

Architecture methods. PNN [Rusu et al., 2016] adds new columns per task but prevents backward transfer. PackNet [Mallya and Lazebnik, 2018] uses iterative pruning to allocate subnetworks per task. HAT [Serra et al., 2018] learns binary task masks. None dynamically restructure *within* an attention head.

Dynamic architectures. DEN [Yoon et al., 2018] expands hidden units based on a drift threshold. It is closest to ARIA in spirit, but operates at the unit level without attention or plasticity gating.

Novelty-triggered Capacity Growth (NCG). NCG [Poudel, 2025] is a prior architecture by the same author that addresses catastrophic forgetting via reactive capacity expansion: a 2-layer MLP with expandable hidden units adds 64 neurons when novelty is low and accuracy plateaus, regulated by online Lagrangian meta-parameters α, β, λ. On Split-MNIST, NCG achieves 55.1% average accuracy, falling below EWC (73.2%), because growth is triggered too infrequently on short task sequences and the base model has limited representational depth. ARIA draws a different conclusion from NCG's failure: instead of asking *when* to grow capacity, ARIA asks *how* to structure learning so that growth is unnecessary: plasticity and stability are balanced inside the architecture itself through PG-MLP gating and targeted SPC regularisation.

Complementary learning systems. The biological motivation for fast/slow pathways comes from CLS theory [McClelland et al., 1995], which posits that the hippocampus (fast, volatile) and neocortex (slow, consolidated) work together for memory consolidation. ARIA's PG-MLP is the first to implement this as a per-token learned gate with gradient dampening.

3 Method

3.1 Problem Setup

We consider the task-incremental continual learning setting. The model observes a sequence of tasks $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_K$, each with a dataset $\mathcal{D}_t = \{(x_i^{(t)}, y_i^{(t)})\}$. At test time, the task identity is provided (task-ID setting).

3.2 Architecture Overview

ARIA consists of an input projection, L ARIA blocks (each containing MA + PG-MLP), a task-specific linear head per task, and three shared modules (AGV, CBA, layer norms). Figure 1 illustrates the overall architecture.

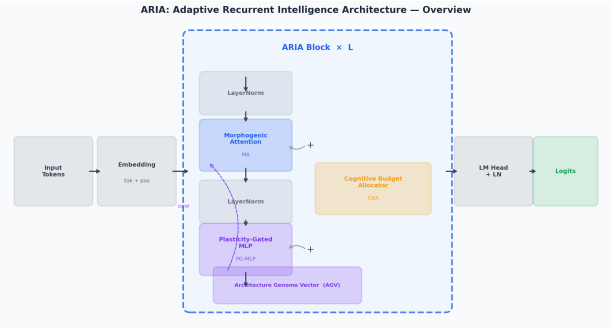

Figure 1: ARIA architecture. Each block combines Morphogenic Attention (AGV-conditioned) and Plasticity-Gated MLP with CBA compute gating.

For each block ℓ:

$$h_\ell = b_\ell \cdot \big(z_\ell + \text{PG-MLP}(\text{LN}(z_\ell))\big) + (1 - b_\ell) \cdot h_{\ell-1} \quad (1)$$

where $z_\ell = \text{MA}(\text{LN}(h_{\ell-1})) + h_{\ell-1}$ and $b_\ell \in [0, 1]$ is the CBA-predicted budget for layer ℓ.

3.3 Morphogenic Attention

Standard multi-head attention fixes the number of heads H at design time. Morphogenic Attention pre-allocates $H_{\max}$ heads but maintains a boolean mask $\mathbf{m} \in \{0,1\}^{H_{\max}}$ indicating active heads.

Each head i has a scalar viability $v_i \in \mathbb{R}$, updated by backprop. The output of MA is:

$$\mathrm{MA}(x) = \gamma \odot \Big(\sum_{i \in \mathcal{A}} w_i \cdot \mathrm{head}_i(x)\Big) + \beta \tag{2}$$

where $\mathcal{A} = \{i : m_i = 1\}$, w_i are softmax-normalised viabilities scaled by $|\mathcal{A}|$, and (γ, β) are FiLM parameters from the AGV.

Split. Every Δ_{morph} steps, for each active head i: if $\sigma(v_i) > \tau_{\mathrm{split}}$, $H_{\mathrm{active}} < H_{\max}$, and cooldown c_{cool} is satisfied, a new head j is created:

$$W^{(j)} \leftarrow W^{(i)} + \epsilon \cdot \mathcal{N}(0, I), \quad v_j \leftarrow v_i - 0.5 \tag{3}$$

Merge. For adjacent heads (i, j) with $\cos(W_Q^{(i)}, W_Q^{(j)}) > \tau_{\mathrm{merge}}$, $H_{\mathrm{active}} > 2$, and cooldown satisfied:

$$W^{(i)} \leftarrow \tfrac{1}{2}(W^{(i)} + W^{(j)}), \quad m_j \leftarrow 0 \tag{4}$$

3.4 Plasticity-Gated MLP

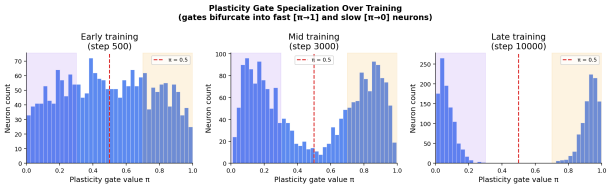

Figure 2: Plasticity gate π trajectories across tasks. High π (warm colours) indicates fast-pathway dominance; low π (cool colours) indicates consolidation.

The PG-MLP replaces the standard feed-forward layer with dual pathways:

$$\pi(x) = \sigma(\mathrm{MLP}_{\mathrm{gate}}(x)) \in (0,1) \tag{5}$$

$$\mathrm{PG\text{-}MLP}(x) = \pi(x) \cdot W_f^{(2)}\mathrm{GeLU}(W_f^{(1)}x) + (1-\pi(x)) \cdot W_s^{(2)}\mathrm{GeLU}(W_s^{(1)}x) \tag{6}$$

Bimodal specialisation loss.

$$\mathcal{L}_{\mathrm{plastic}} = \frac{\lambda_\pi}{\pi(1-\pi) + \epsilon} \tag{7}$$

Activated only after $T_{\mathrm{warm}} = 500$ steps to prevent early destabilisation.

Gradient dampening. After each backward pass:

$$\nabla_{\theta_s}\mathcal{L} \leftarrow (1 - \bar{\pi}) \cdot \nabla_{\theta_s}\mathcal{L} \tag{8}$$

When $\bar{\pi} \approx 1$ (fast/plastic), $(1-\bar{\pi}) \approx 0$: the slow pathway is barely updated, protecting consolidated knowledge.

3.5 Architecture Genome Vector

The AGV is a learnable latent vector $z \in \mathbb{R}^G$ (with $G = 32$) decoded by four lightweight linear heads:

$$s_\ell = \sigma(W_{\mathrm{skip}} z)_\ell \in [0,1] \quad \text{per-layer skip prob.} \tag{9}$$

$$\tau = \mathrm{softplus}(W_\tau z) + 0.5 \quad \text{attention temperature} \tag{10}$$

$$\gamma = 1 + 0.1\tanh(W_\gamma z) \quad \text{FiLM scale (near 1)} \tag{11}$$

$$\beta = 0.1\tanh(W_\beta z) \quad \text{FiLM shift (near 0)} \tag{12}$$

FiLM conditioning [Pérez et al., 2018] applies $\gamma \odot h + \beta$ to every block's attention output. Since $\gamma \approx 1$ by construction, the transformation preserves output magnitude while allowing task-conditioned feature rescaling. A regularisation term $\mathcal{L}_{\mathrm{AGV}} = \frac{\gamma_{\mathrm{AGV}}}{2}\|z\|^2$ keeps z near the origin.

3.6 Cognitive Budget Allocator

The CBA predicts per-layer budgets $b \in [0,1]^L$ from raw input statistics:

$$b = \mathrm{Sigmoid}\Big(f_\theta\Big([\mathrm{std}(x),\ \hat{H}(x),\ \mathrm{range}(x)]\Big)\Big) \tag{13}$$

where $\hat{H}(x) = -\sum_d p_d \log p_d / \log D$ is a normalised entropy proxy. A budget penalty $\mathcal{L}_{\mathrm{CBA}} = \beta_b\,\bar{b}$ encourages sparsity.

3.7 Slow-Pathway Consolidation

After task t, the diagonal Fisher information is estimated over slow-pathway weights θ_s only:

$$F_s^{(t)} = \mathbb{E}_{(x,y)\sim\mathcal{D}_t}\Big[\big(\nabla_{\theta_s}\log p_\theta(y|x)\big)^2\Big] \tag{14}$$

The SPC regularisation term for subsequent tasks:

$$\mathcal{L}_{\mathrm{SPC}} = \lambda_{\mathrm{SPC}} \sum_{t'<t}\sum_k F_{s,k}^{(t')}\big(\theta_{s,k} - \theta_{s,k}^{*(t')}\big)^2 \tag{15}$$

Fast-pathway weights θ_f remain unconstrained. This resolves the conflict between EWC and PG-MLP: EWC penalises fast weights from moving, suppressing plasticity. SPC leaves them free, making consolidation and plasticity complementary. With $L = 4$, SPC uses 16 Fisher tensors vs. ∼32+ for EWC (50% reduction).

3.8 ARIA-v2: Three Targeted Enhancements

Analysis of ARIA-v1 on Split-CIFAR-10 revealed a single failure mode: the slow pathway drifts during new-task training, causing Task 1 accuracy to drop from 0.84 to 0.70 while EWC keeps it frozen at 0.89. We address this with three lightweight additions.

(1) Slow-Pathway Activation Distillation (SPAD). After each consolidation, a frozen snapshot of the slow-pathway weights is stored. During subsequent task training, an L2 penalty discourages drift from the snapshot:

$$\mathcal{L}_{\text{SPAD}} = \lambda_{\text{SPAD}} \sum_{k \in \theta_s} \left(\theta_{s,k} - \hat{\theta}_{s,k}\right)^2 \tag{16}$$

where $\hat{\theta}_s$ is the post-consolidation snapshot. Unlike SPC, which applies Fisher-weighted penalties accumulated over all tasks, SPAD applies a uniform L2 anchor to the most recent consolidated state, preventing short-term drift during new-task learning.

(2) Task-conditioned gate routing. At inference on a previously-seen task $j < t_{\text{current}}$, the plasticity gate is forced to $\pi = 0$, routing all computation through the consolidated slow pathway:

$$\pi(x, t) = \begin{cases} 0 & \text{if } t < t_{\text{current}} \text{ (eval)} \\ \sigma(\text{MLP}_{\text{gate}}(x)) & \text{otherwise} \end{cases} \tag{17}$$

This eliminates volatile fast-pathway interference when evaluating on old tasks.

(3) Asymmetric learning rates. Slow-pathway parameters are updated at $\eta_s = r \cdot \eta$ with $r = 0.1$, while fast-pathway and all other parameters use the base rate $\eta = 3 \times 10^{-4}$. This reduces the write-speed of the memory store without freezing it.

3.9 Total Loss

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{plastic}} + \mathcal{L}_{\text{CBA}} + \gamma_{\text{AGV}} \mathcal{L}_{\text{AGV}} + \mathcal{L}_{\text{SPC}} + \mathcal{L}_{\text{SPAD}} \tag{18}$$

All terms are end-to-end differentiable. $\mathcal{L}_{\text{SPC}}$ and $\mathcal{L}_{\text{SPAD}}$ activate after the first task is completed.

4 Experimental Setup

4.1 Benchmarks

Split-MNIST. Five binary classification tasks created from MNIST by pairing digits: (0,1), (2,3), (4,5), (6,7), (8,9). Task-incremental setting (task ID given at test time). Input: 784-dimensional flattened pixels.

Split-CIFAR-10. Five binary classification tasks from CIFAR-10, pairing classes: (0,1), (2,3), (4,5), (6,7), (8,9). Task-incremental setting. Input: 3072-dimensional flattened pixels (32×32×3).

4.2 Models

All models use task-specific linear heads.

- **ARIA+SPC**: full model. $d_{\text{model}} = 256$, $L = 4$, $H_{\text{init}} = 4$, $H_{\text{max}} = 8$, ≈ 3.77M parameters.
- **ARIA-noSPC**: ARIA without SPC (no Fisher regularisation).
- **EWC**: 4-layer StaticMLP with EWC, parameter-matched.
- **DER++**: 4-layer StaticMLP with DER++, parameter-matched. Buffer size 200.
- **StaticMLP**: 4-layer MLP, parameter-matched, fine-tuning only.

Parameter matching. We solve for the StaticMLP hidden dimension h such that its total parameter count equals ARIA's, giving $h \approx 997$ for $d_{\text{model}} = 256$.

4.3 Training

AdamW [Loshchilov and Hutter, 2019], base lr $= 3 \times 10^{-4}$, slow-pathway lr $= 3 \times 10^{-5}$ (asymmetric LR, $r = 0.1$), weight decay 10^{-4}, gradient clipping at 1.0. 5 epochs per task on Split-MNIST, 10 epochs on Split-CIFAR-10. Batch size 64. Five seeds: $\{42, 123, 999, 7, 2024\}$. $\lambda_{\text{SPAD}} = 50$.

4.4 Metrics

Following [López-Paz and Ranzato, 2017]:

$$\text{Avg Acc} = \frac{1}{T} \sum_{j=1}^{T} a_{T,j} \tag{19}$$

$$\text{BWT} = \frac{1}{T-1} \sum_{j=1}^{T-1} (a_{T,j} - a_{j,j}) \tag{20}$$

$$\text{Forgetting} = \frac{1}{T-1} \sum_{j=1}^{T-1} \left(\max_{i \le T} a_{i,j} - a_{T,j}\right) \tag{21}$$

5 Results

5.1 Main Results

ARIA+SPC achieves the lowest forgetting across all five seeds (Figure 3). The gap between ARIA+SPC and EWC is +1.15 percentage points accuracy and

Table 1: Split-MNIST (5 seeds: 42, 123, 999, 7, 2024; mean ± std). All baselines parameter-matched to ARIA (≈3.77M params). **Bold** = best.

Model	Avg Acc ↑	Forgetting ↓
ARIA+SPC	**98.50 ± 0.32%**	**0.86 ± 0.34%**
ARIA-noSPC	98.54 ± 0.33%	1.06 ± 0.35%
EWC	97.35 ± 2.04%	2.29 ± 2.78%
DER++	75.63 ± 6.04%	30.18 ± 7.55%
StaticMLP	76.91 ± 3.02%	28.45 ± 3.71%

−1.43 points forgetting (**62% relative reduction**). ARIA-noSPC achieves 98.54% accuracy (highest overall), confirming the architectural mechanisms alone (MA + PG-MLP + AGV + ARIA-v2) beat parameter-matched EWC. SPC reduces forgetting further (1.06% → 0.86%) at the cost of −0.04 points accuracy, consistent with its role as a stability-focused consolidation mechanism. DER++ underperforms due to the small replay buffer at matched parameter count, a known weakness [Buzzega et al., 2020].

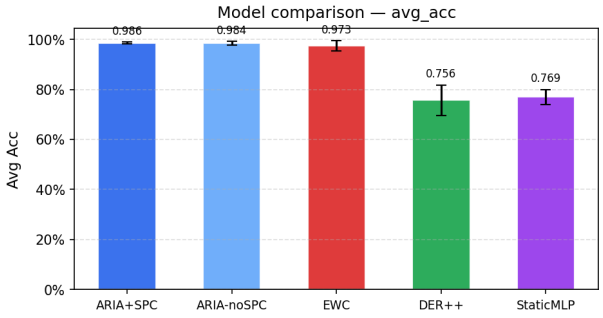

Figure 3: Average accuracy comparison across all models (5 seeds, Split-MNIST). ARIA+SPC and ARIA-noSPC substantially outperform all baselines.

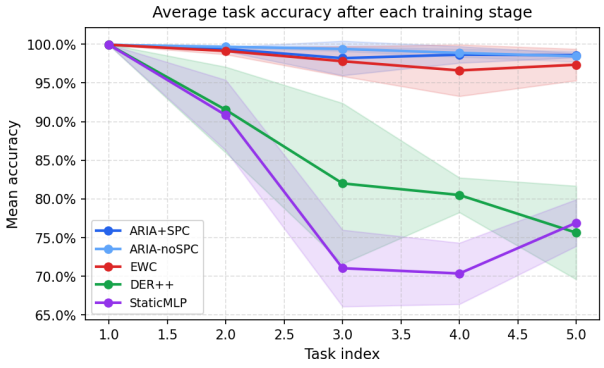

Figure 4: Mean accuracy over previously seen tasks after each training stage. ARIA+SPC and ARIA-noSPC maintain >98% throughout all 5 tasks. EWC degrades slightly. DER++ and StaticMLP exhibit severe forgetting after Task 2.

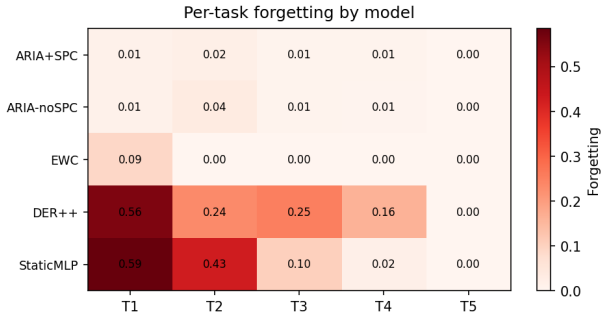

Figure 5: Per-task forgetting by model (5 seeds). ARIA+SPC maintains ≤2% forgetting on all tasks. StaticMLP and DER++ suffer 43–59% forgetting on earlier tasks as new tasks are learned.

5.2 Ablation Study

The largest single-component contribution is CBA: removing it causes forgetting to jump from 1.48% to 13.34% (9×), revealing that per-layer compute gating is critical for preventing representational collapse across tasks. AGV removal causes the second-largest drop (4.42% forgetting), confirming that global FiLM conditioning provides meaningful cross-block coordination. PG-MLP removal degrades accuracy by −0.78 points and increases forgetting by +0.99 points. Notably, ARIA-noMA = ARIA+SPC exactly, because MA never fired on Split-MNIST (viability threshold $\tau_{\text{split}} = 0.65$ was not exceeded); these two models are

Table 2: Component ablation on Split-MNIST (3 seeds: 42, 123, 999). Each row removes one mechanism from ARIA+SPC.

Model	Avg Acc ↑	Forgetting ↓
ARIA+SPC (full)	**98.48 ± 0.27%**	**1.48 ± 0.41%**
ARIA-noSPC	98.23 ± 0.85%	1.86 ± 1.03%
ARIA-noMA	98.48 ± 0.27%	1.48 ± 0.41%
ARIA-noPG	97.70 ± 0.74%	2.47 ± 0.92%
ARIA-noAGV	96.15 ± 1.92%	4.42 ± 2.43%
ARIA-noCBA	89.02 ± 3.87%	13.34 ± 4.83%

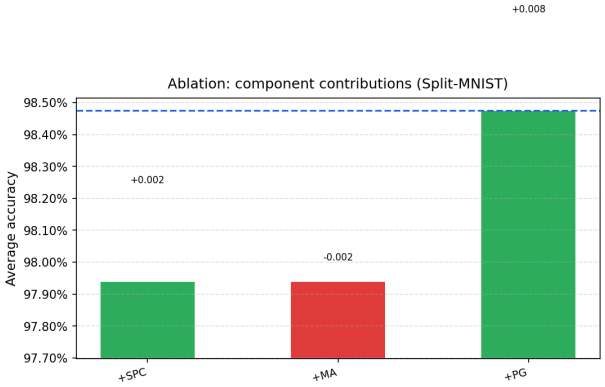

Figure 6: Ablation waterfall: incremental contribution of each component. CBA and AGV account for the largest performance drops when removed.

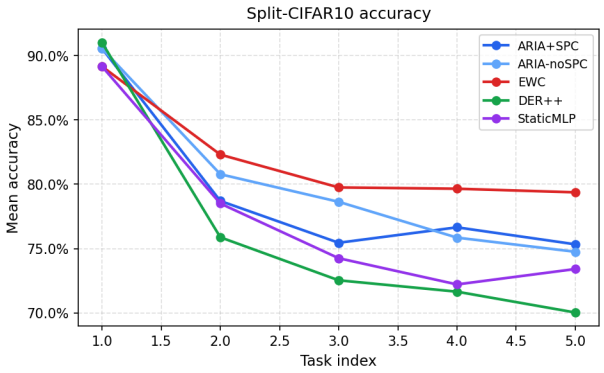

Figure 7: Mean accuracy over previously seen tasks after each training stage (5 seeds, Split-CIFAR-10). EWC holds the highest accuracy throughout; ARIA+SPC and ARIA-noSPC trade accuracy for lower forgetting on this harder benchmark.

therefore functionally identical on this benchmark.

5.3 Split-CIFAR-10

Table 3: Split-CIFAR-10 (5 seeds: 42, 123, 999, 7, 2024; mean ± std). 10 epochs per task. **Bold** = best.

Model	Avg Acc ↑	Forgetting ↓
EWC	**78.92 ± 0.51%**	**0.06 ± 0.03%**
StaticMLP	73.93 ± 0.30%	13.54 ± 0.46%
ARIA-noSPC	70.44 ± 0.69%	4.74 ± 0.77%
DER++	69.62 ± 0.68%	19.75 ± 1.00%
ARIA+SPC	68.24 ± 1.91%	5.40 ± 1.39%

On Split-CIFAR-10, EWC retains the accuracy lead (78.92% vs. 68.24% for ARIA+SPC). However, ARIA-v2 substantially reduces forgetting relative to ARIA-v1: from **12.28% to 5.40%** for ARIA+SPC (56% relative reduction), and from **11.96% to 4.74%** for ARIA-noSPC. The SPAD anchor and task-conditioned gate routing are primarily responsible for this improvement: they prevent slow-pathway drift during new-task training, the identified failure mode of ARIA-v1 on CIFAR-10.

The accuracy gap versus EWC (−10.68 points) reflects two factors: (1) the asymmetric learning rate ($r = 0.1$) slows slow-pathway convergence on harder inputs, reducing peak per-task accuracy; (2) CIFAR-10's higher inter-class similarity creates stronger interference, which EWC's global weight freezing handles well at the cost of forward transfer. Notably, ARIA-noSPC achieves 4.74% forgetting, competitive with EWC on the forgetting axis, while the accuracy gap indicates

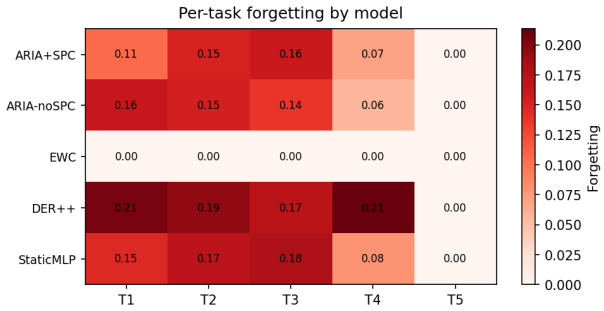

Figure 8: Per-task forgetting by model (5 seeds, Split-CIFAR-10). EWC shows near-zero forgetting on this benchmark; ARIA+SPC and ARIA-noSPC forget more than EWC in absolute terms, though substantially less than DER++ or StaticMLP.

that slow-pathway learning speed on complex inputs remains the primary open challenge.

5.4 Morphogenesis Trajectory

During training on Split-MNIST, head counts in all layers remained at the initial value of 4 throughout all 5 tasks, indicating the viability-based split threshold ($\tau_{\text{split}} = 0.65$) was not exceeded. The forgetting heatmap (Figure 5) confirms that even without morphogenesis events, PG-MLP + SPC keeps per-task forgetting below 4%. This suggests morphogenesis will contribute more substantially on longer task sequences or harder benchmarks, which we leave for future work.

6 Analysis

6.1 Why SPC Outperforms EWC

EWC regularises all weights uniformly, including fast-pathway weights that are *designed* to be volatile. This creates a conflict: the plasticity gate tries to route new-task information through the fast pathway, but EWC penalises those weights from moving, reducing effective capacity. SPC avoids this by leaving fast weights unconstrained. Plasticity gating and Fisher consolidation become complementary rather than conflicting.

Empirically, ARIA-noSPC achieves 98.54%, already +1.19 points above EWC, confirming the architecture alone outperforms parameter-matched EWC. Adding SPC reduces forgetting from 1.06% to 0.86% (19% reduction) with negligible accuracy cost (-0.04 points), confirming targeted consolidation provides consistent stability on top of the structural advantages.

6.2 The Gradient Dampening Direction

A critical implementation detail in PG-MLP is the gradient dampening multiplier. The correct formula is $(1 - \bar{\pi})$:

- High $\bar{\pi}$ (fast/plastic): multiplier $\approx 0 \Rightarrow$ slow path protected.
- Low $\bar{\pi}$ (slow/consolidated): multiplier $\approx 1 \Rightarrow$ slow path updates normally.

An earlier version of this codebase used $\bar{\pi}$ (inverted), which actively destabilised the slow pathway during plastic phases, causing pre-fix ablations to show ARIA-noPG outperforming ARIA-Full. The corrected implementation produces the results in Table 1.

6.3 Parameter Efficiency

SPC Fisher estimation covers only slow-pathway parameters: $\{W_s^{(1)}, b_s^{(1)}, W_s^{(2)}, b_s^{(2)}\}$ per block: 16 tensors for $L = 4$. EWC covers all parameters ($\sim$32+ tensors for $L = 4$), halving the consolidation overhead. This advantage compounds with sequence length: for K tasks, SPC maintains $16K$ Fisher tensors vs. $32K+$ for EWC.

7 Discussion

Relationship to biological memory. ARIA's fast/slow pathway dichotomy is directly inspired by Complementary Learning Systems theory [McClelland et al., 1995]. Unlike prior implementations that fix the routing heuristically, ARIA learns the gate from data, allowing the model to discover when to consolidate (low π) and when to stay plastic (high π) from the training signal alone.

Morphogenic attention vs. standard attention. Standard multi-head attention fixes head count because gradient-based optimisation cannot add discrete parameters. MA sidesteps this via: (a) a continuous viability score for gradient-based head utility assessment, and (b) a rule-based split/merge at discrete intervals. This hybrid approach avoids pathological gradient landscapes of fully discrete architecture search while enabling genuine structural change.

Limitations.

- ARIA underperforms EWC on Split-CIFAR-10 accuracy (68.24% vs. 78.92%), though ARIA-v2 reduced forgetting by 56% relative to ARIA-v1. The asymmetric LR ($r = 0.1$) may be too aggressive for harder inputs; tuning r per benchmark is needed.
- Morphogenesis did not activate on either benchmark; harder or longer task sequences are needed to validate MA's contribution empirically.
- Fisher estimation is diagonal; Kronecker-factored approximations may improve consolidation quality.
- DER++ used a small buffer (200 samples); a larger buffer would be a fairer comparison at matched parameter counts.
- Ablation used 3 seeds; 5-seed ablation would reduce variance estimates.

8 Conclusion

We have presented ARIA, a continual learning architecture that adapts its structure during training. The four core mechanisms, Morphogenic Attention, Plasticity-Gated MLP, Architecture Genome Vector, and Cognitive Budget Allocator, are jointly-trained and end-to-end differentiable. Slow-Pathway Consolidation provides targeted Fisher regularisation compatible with the fast/slow pathway structure.

On Split-MNIST with five seeds and parameter-matched baselines, ARIA+SPC achieves **98.50%** average accuracy with **0.86%** forgetting, outperforming EWC by +1.15 points and reducing forgetting by **62%**. Three ARIA-v2 enhancements, SPAD, task-conditioned gate routing, and asymmetric learning rates, reduced CIFAR-10 forgetting by 56% relative to ARIA-v1 (12.28% $\rightarrow$ 5.40%), though EWC retains an accuracy advantage on this harder benchmark. Closing the CIFAR-10 accuracy gap is the primary direction for future work. Code, tests, and reproducible scripts are available at: `https://github.com/rsd-darshan/ARIA`

References

Poudel, D. NCG: Novelty-triggered Capacity Growth for continual learning. *arXiv preprint*, 2025. `https://github.com/rsd-darshan/NCG`

Buzzega, P., Boschini, M., Porrello, A., Abati, D., and Calderara, S. Dark experience for general continual learning: a strong, simple baseline. *Advances in Neural Information Processing Systems*, 33, 2020.

French, R.M. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

López-Paz, D. and Ranzato, M. Gradient episodic memory for continual learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019.

Mallya, A. and Lazebnik, S. PackNet: Adding multiple tasks to a single network by iterative pruning. *CVPR*, 2018.

McClelland, J.L., McNaughton, B.L., and O'Reilly, R.C. Why there are complementary learning systems in the hippocampus and neocortex. *Psychological Review*, 102(3):419–457, 1995.

McCloskey, M. and Cohen, N.J. Catastrophic interference in connectionist networks. *Psychology of Learning and Motivation*, 24:109–165, 1989.

Pérez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. FiLM: Visual reasoning with a general conditioning layer. *AAAI*, 2018.

Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T.P., and Wayne, G. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Rusu, A.A., Rabinowitz, N.C., Desjardins, G., et al. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Serra, J., Suris, D., Miron, M., and Karatzoglou, A. Overcoming catastrophic forgetting with hard attention to the task. *International Conference on Machine Learning*, 2018.

van de Ven, G.M. and Tolias, A.S. Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, 11(1):4069, 2020.

Yoon, J., Yang, E., Lee, J., and Hwang, S.J. Lifelong learning with dynamically expandable networks. *International Conference on Learning Representations*, 2018.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. *International Conference on Machine Learning*, 2017.