# Distil and Evolve: Compact Adaptive Document Classification with Continual Learning for Low-Resource Settings

**Mehdi Soufiane**[1,2,3] **Mourad Hammale**[1] **Kawtar Zerhouni**[1]
**Adil Ahidar-Coutrix**[1] **Rachid Benouini**[2,3]
[1]Laboratory of Complex Systems, École Centrale Casablanca, Morocco
[2]LEYTON, Casablanca, Morocco [3]LEYTON CognitX
{mehdi.soufiane, mourad.hammale}@centrale-casablanca.ma
{adil.ahidar, kawtar.zerhouni}@centrale-casablanca.ma

## Abstract

Document image classification is the entry point of document-processing pipelines, yet the institutions that need them most operate under tight compute and annotation budgets. Knowledge distillation (KD) compresses large document Transformers for such deployment, but whether a heavily distilled model can still undergo continual learning (CL) without catastrophic forgetting, and which CL strategy suits it, is unclear. We distil a DiT-Large teacher (307 M parameters) into a ViT-Tiny student (5.5 M, 21.2 MiB) and evaluate four CL methods in a two-phase class-split protocol on RVL-CDIP across five random seeds. Replay-based methods significantly outperform regularisation-based ones: DER++ retains $90.4 \pm 1.2\%$ of prior accuracy against $75.9 \pm 2.1\%$ for the best regulariser, a 14.5 pp paired difference ($p < 10^{-4}$). Every pairwise ordering is significant and holds at every seed. Because hyperparameters were selected on one seed, we also measure the resulting optimism: the selection seed overstates retention by up to 7.4 pp, and most for the weakest method. Per-class analysis shows macro retention conceals concentrated damage, with two of five target classes absorbing most of the loss. We profile each strategy against deployment metrics, including instrumented GPU energy and single-thread CPU latency, and find that counting method preparation shrinks the time advantage of regularisation over replay from 9.1 to 3.3 minutes. A same-size non-distilled control starts near chance and therefore cannot isolate a distillation effect; we report it as a diagnostic only. We release the distilled model, the harness, and all seed-level artifacts.

## 1 Introduction

Automated classification of document images, including invoices, contracts, forms, and scientific reports, is a core component of enterprise information systems. Transformer-based document understanding models such as DiT (Li et al., 2022) and LayoutLMv3 (Huang et al., 2022) achieve strong accuracy but contain hundreds of millions of parameters, making them impractical for latency-sensitive or resource-constrained deployment. This barrier is most acute in low-resource contexts, where digitisation pipelines must run on limited infrastructure and large annotated corpora are unavailable. Document image classification is largely script-agnostic, operating on visual layout rather than textual content, so compact adaptive classifiers are a transferable building block for document processing in low-resource languages, including the multilingual administrative archives common in North African institutions. Knowledge distillation (KD) (Hinton et al., 2015) addresses the compute side of this gap by transferring a large teacher's representations into a compact student, with modest accuracy loss (Gou et al., 2021).

Compression is only part of the deployment challenge. Once in production, a document classifier must adapt as organisational needs evolve. Naïvely fine-tuning on new classes causes catastrophic forgetting (McCloskey and Cohen, 1989; French, 1999): the model loses previously learned classes. Continual learning (CL) methods (Wang et al., 2024) mitigate this, but have been studied primarily on full-capacity models. A distilled student has substantially reduced representational redundancy compared with its teacher: fewer layers, a narrower hidden dimension, fewer attention heads. This motivates a practical question. Parameter-regularisation methods rely on spare weight-space directions where new knowledge can be absorbed without disturbing old representations, and a compact model offers fewer of them; replay methods instead inject direct gradient signal for past tasks at every step. Which family should a practitioner deploying a distilled doc-



*Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages*, pages 131–139
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_015

ument classifier actually choose?

We stress at the outset what this study can and cannot settle. Our experiments hold capacity fixed at ViT-Tiny. They therefore establish which family wins *on a compact distilled document classifier*, with quantified uncertainty, but they cannot attribute that outcome to compactness: separating "regularisation fails because the model is small" from "regularisation underperforms replay here as it commonly does at full capacity" requires a capacity sweep that we did not run. We treat the capacity argument as motivation, not as a finding.

Two families offer contrasting mechanisms. Regularisation-based methods penalise changes to parameters deemed important for previous tasks. Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) measures importance via the diagonal Fisher information matrix, while Memory Aware Synapses (MAS) (Aljundi et al., 2018) uses the sensitivity of the network's output norm to parameter perturbations. Related approaches include Synaptic Intelligence (Zenke et al., 2017) and Learning without Forgetting (LwF) (Li and Hoiem, 2017), which constrains the output distribution rather than the parameters. Replay-based methods store and re-expose past examples. Experience Replay (ER) (Chaudhry et al., 2019) replays raw samples from a memory buffer, while Dark Experience Replay++ (DER++) (Buzzega et al., 2020) additionally preserves the previous model's logits for buffered samples. Among replay variants, iCaRL (Rebuffi et al., 2017) and GEM (Lopez-Paz and Ranzato, 2017) further constrain gradient updates.

Our contributions are:

- a five-seed evaluation of four CL methods on a distilled ViT-Tiny document classifier, with paired significance tests, establishing that the replay advantage is a stable ordering rather than a single-run artifact;
- a direct measurement of how much single-seed hyperparameter selection inflates reported retention, and of which methods it flatters most;
- a per-class analysis showing that macro retention hides concentrated degradation, and a corrected reading of the non-distilled control, which we show cannot support a distillation claim;
- an instrumented deployment profile including GPU energy, method preparation cost, and single-thread CPU latency, released with all seed-level artifacts.

## 2 Related Work

**Document representations.** LayoutLM introduced joint text-and-layout pre-training (Xu et al., 2020); LayoutLMv2 extended it multimodally (Xu et al., 2021); LayoutLMv3 unifies text and image masking (Huang et al., 2022); DiT studies self-supervised pre-training directly on document images (Li et al., 2022). These concern representation learning and backbone design. We instead fix one compact student and study its post-deployment adaptation.

**Continual learning.** Wang et al. (2024) survey CL across five families but do not address how compression modulates their relative effectiveness. CL has been studied for general image recognition (Rebuffi et al., 2017) and for language (d'Autume et al., 2019), yet CL on distilled document classifiers remains little explored.

**Replay-free alternatives.** Recent parameter-efficient work offers an important alternative to the four methods compared here. L2P learns a prompt pool over a frozen backbone (Wang et al., 2022), and CODA-Prompt learns input-conditioned prompt components without rehearsal data (Smith et al., 2023). These avoid retaining a document buffer entirely, which matters where page retention is restricted by governance rather than by storage. We did not implement them: both assume a frozen, high-capacity pre-trained backbone, and their behaviour on a 5.5 M-parameter student that must itself keep adapting is an open question we flag as the most relevant extension to our comparison.

## 3 Methodology

### 3.1 Knowledge Distillation Pipeline

We employ offline logit-based distillation on RVL-CDIP (Harley et al., 2015), a benchmark of 16 document classes. The original dataset contains approximately 400,000 images; our experiments use a curated subset[1] of 30,400 training and 3,200 test images. The subset was adopted for tractability under our compute budget; its sampling proce-

[1] `sitloboi2012/rvl_cdip_large_dataset` on Hugging Face Datasets, pinned at revision `05a84f4e`.

dure is not documented by the source, so our absolute accuracies are not directly comparable to published full-split RVL-CDIP numbers, and we do not claim they are. All comparisons in this paper are internal to the subset. The student minimises

$$\mathcal{L}_{\text{KD}} = \alpha \cdot \mathcal{L}_{\text{CE}}(y, \sigma(z_S)) + (1-\alpha) \cdot T^2 \cdot \text{KL}(\sigma(z_T/T) \| \sigma(z_S/T)) \quad (1)$$

where $z_T$ and $z_S$ are teacher and student logits, $T$ is the temperature, $\alpha$ balances hard and soft targets, and $\sigma$ is softmax.

The teacher is DiT-Large (Li et al., 2022) (307 M parameters, 24 layers). The student is ViT-Tiny (5,527,504 parameters, 12 layers, hidden size 192, 3 heads, 21.2 MiB on disk, 1.26 GMACs per image), a 55.5× parameter reduction. The distilled student is the starting point $\theta_{\text{orig}}$ for all CL experiments.

### 3.2 Two-Phase Protocol and Metrics

We evaluate CL using a two-phase protocol (Figure 1). In Phase 1 (Task A, specialisation), $\theta_{\text{orig}}$ is fine-tuned without CL protection on 5 target classes, producing a specialist $\theta_{\text{naive}}$. In Phase 2 (Task B), each CL method starts from $\theta_{\text{naive}}$ and fine-tunes on the 11 remaining classes. All CL anchors are computed from $\theta_{\text{naive}}$ on Task-A data only. We also evaluate a No-CL Phase-2 condition as a lower bound.

Because $\theta_{\text{orig}}$ already predicts all 16 classes, Task B is a *recovery* group rather than a set of globally unseen classes. We therefore use "two-phase class-split adaptation" rather than "class-incremental learning". This matches a deployed classifier that is temporarily specialised and later broadened; it does not cover the case where genuinely new labels arrive after deployment.

Let $a_c(\theta)$ be the per-class accuracy of $\theta$, $\mathcal{T}_A$ the 5 target classes and $\mathcal{T}_B$ the 11 remaining. Task-A retention measures the fraction of $\theta_{\text{naive}}$'s target-class accuracy preserved:

$$\text{Ret} = \frac{1}{|\mathcal{T}_A|} \sum_{c \in \mathcal{T}_A} \frac{a_c(\theta_{\text{CL}})}{a_c(\theta_{\text{naive}})} \quad (2)$$

Task-B efficiency measures recovery of $\theta_{\text{orig}}$'s accuracy on the remaining classes:

$$\text{Eff} = \frac{\frac{1}{|\mathcal{T}_B|} \sum_{c \in \mathcal{T}_B} a_c(\theta_{\text{CL}})}{\frac{1}{|\mathcal{T}_B|} \sum_{c \in \mathcal{T}_B} a_c(\theta_{\text{orig}})} \quad (3)$$

Retention is a mean of per-class ratios, so values above 100% are possible when a class improves. We also report global 16-class accuracy, because retention alone rewards a method that simply resists learning Task B.

We complement these with deployment metrics recorded by the instrumented pipeline: parameter count, on-disk size, inference latency, wall-clock adaptation time separated into method preparation and training, peak GPU memory, GPU energy, and auxiliary adaptation storage $S_{\text{aux}}$, the persistent memory a method needs beyond the weights (importance and anchor matrices, or the replay buffer).

### 3.3 Continual Learning Methods

EWC (Kirkpatrick et al., 2017) adds $\frac{\lambda}{2} \sum_i F_i(\theta_i - \theta_{\text{naive},i})^2$ to the loss, with $F_i$ the Fisher diagonal computed on $\theta_{\text{naive}}$ from Task-A samples. MAS (Aljundi et al., 2018) shares this quadratic form but estimates importance as the mean absolute gradient of the output $L_2$ norm, an output-based rather than loss-curvature signal. ER (Chaudhry et al., 2019) maintains a class-balanced buffer of Task-A samples and adds a replay cross-entropy term at each Phase-2 step. DER++ (Buzzega et al., 2020) additionally stores $\theta_{\text{naive}}$'s logits for buffered samples and adds an MSE term constraining outputs to stay close to them. Note that ER weights replay cross-entropy by 1 whereas DER++ weights it by $\beta$ and adds $\alpha$-weighted logit MSE, so their comparison reflects the combined configuration rather than isolating the stored-logit term.

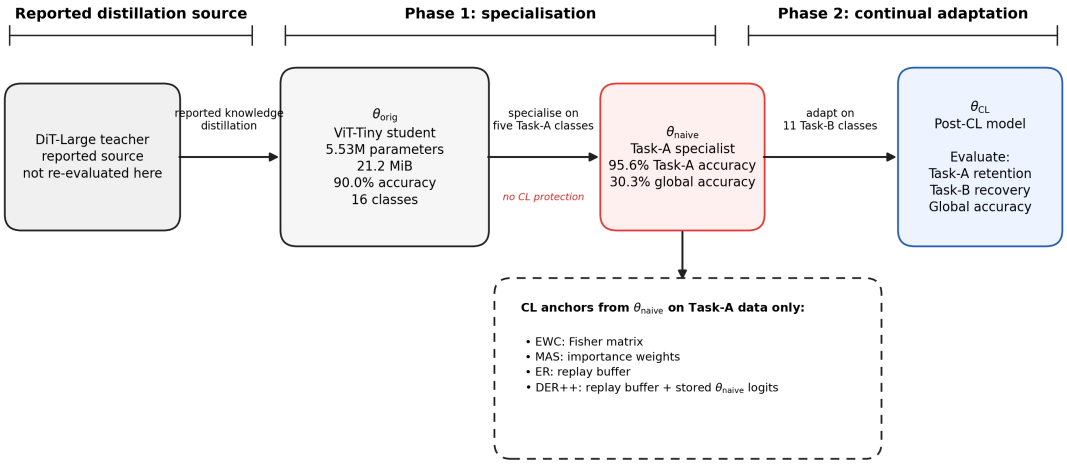


Figure 1: Two-phase class-split continual-learning protocol. The distilled ViT-Tiny student $\theta_{\text{orig}}$ (5.5 M parameters, obtained by KD from DiT-Large, 307 M) is the single starting point. Phase 1 fine-tunes it without CL protection on 5 target classes (Task A), producing the specialist $\theta_{\text{naive}}$. Phase 2 starts from $\theta_{\text{naive}}$ and fine-tunes on the remaining 11 classes (Task B) under a CL method. All anchors (Fisher, importance weights, replay buffers, stored logits) derive from $\theta_{\text{naive}}$ on Task-A data only. The whole protocol is repeated independently for each of five seeds.

## 4 Experimental Setup

We use the 16-class RVL-CDIP subset[1] with 30,400 training and 3,200 test images (200 per class). Task A comprises 5 semantically diverse classes: email (2), handwritten (3), specification (7), invoice (11), presentation (12). The remaining 11 form Task B. All experiments use AdamW with learning rate $5 \times 10^{-5}$, weight decay 0.01, batch size 8, and gradient clipping at 1.0. Phase 1 uses a constant learning rate over 10 epochs; Phase 2 uses cosine annealing with 5% linear warmup over 10 epochs. Each phase draws up to 8,000 training samples. Fisher and importance matrices are computed on all available Task-A samples ($\sim$9,500), which is common to all compared methods but is not a strict online protocol.

**Seeds, and the separation of selection from evaluation.** A grid search was run on seed 42, selecting per method the configuration maximising Task-A retention: EWC $\lambda \in \{5{,}000, 10{,}000, \mathbf{50{,}000}\}$; MAS $\lambda \in \{\mathbf{5}, 10, 20\}$; ER buffer $\in \{200, 500, \mathbf{1{,}000}\}$; DER++ buffer $\in \{200, 500, \mathbf{1{,}000}\}$ with $\alpha = 0.3, \beta = 0.5$. We then *froze* those choices and ran the complete protocol on four further seeds (43–46), which are therefore held-out evaluations of a selection made elsewhere. Unless stated otherwise, reported values are means $\pm$ sample standard deviations over all five seeds, and comparisons between methods are paired across seeds. Section 5.3 uses the held-out seeds to quantify what the single-seed selection cost us.

LwF (Li and Hoiem, 2017) was evaluated in the sweep but excluded from the main comparison: its retention-selected configuration reaches 100.2% retention at only 21.6% Task-B efficiency and 43.3% global accuracy, i.e. it preserves Task A by declining to learn Task B. We report it as evidence that retention alone is an inadequate selection criterion.

## 5 Results

### 5.1 Main Comparison

The distilled student reaches 90.0% accuracy on the 16-class test set against 93.7% for the DiT-Large teacher, a 3.7 pp drop at 55.5$\times$ fewer parameters. After Phase 1, $\theta_{\text{naive}}$ reaches $96.3 \pm 0.5\%$ on Task-A classes but falls to $30.4 \pm 0.2\%$ globally, confirming catastrophic forgetting of Task B. Unprotected Phase-2 fine-tuning retains only $23.5 \pm 5.5\%$ of Task-A accuracy while recovering Task B efficiently ($97.2 \pm 0.8\%$).

All four CL methods improve substantially on that lower bound, and the ordering is identical at every seed (Table 1, Figure 2). DER++ retains $90.4 \pm 1.2\%$, ahead of ER ($84.1 \pm 1.8$), MAS ($75.9 \pm 2.1$) and EWC ($63.3 \pm 3.8$). DER++ also

achieves the highest global accuracy ($85.6 \pm 0.4\%$), while ER is the strongest Task-B learner ($96.5 \pm 0.5\%$). Notably, the standard deviation shrinks monotonically with retention: EWC varies three times as much across seeds as DER++, so the weakest method is also the least predictable.

Table 1: Main results, mean ± sd over five seeds. A-Ret. = Task-A retention; B-Eff. = Task-B recovery efficiency; Global = 16-class accuracy. Best CL result in bold.

| Method | A-Ret.↑ | B-Eff.↑ | Global↑ |
|---|---|---|---|
| $\theta_{\text{orig}}$ (reference) | — | — | 90.0 |
| $\theta_{\text{naive}}$ (Phase 1) | — | — | $30.4 \pm 0.2$ |
| No-CL Phase-2 | $23.5 \pm 5.5$ | **97.2** $\pm 0.8$ | $67.1 \pm 1.8$ |
| EWC ($\lambda$=50k) | $63.3 \pm 3.8$ | $89.5 \pm 1.2$ | $74.3 \pm 1.4$ |
| MAS ($\lambda$=5.0) | $75.9 \pm 2.1$ | $94.3 \pm 0.7$ | $81.0 \pm 0.9$ |
| ER (buffer=1k) | $84.1 \pm 1.8$ | $96.5 \pm 0.5$ | $84.9 \pm 0.7$ |
| DER++ (buffer=1k) | **90.4** $\pm 1.2$ | $94.6 \pm 0.4$ | **85.6** $\pm 0.4$ |

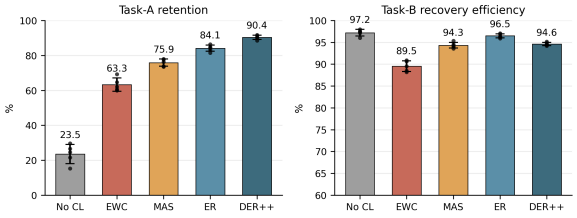


Figure 2: Task-A retention and Task-B recovery efficiency, mean ± sd over five seeds. Individual seed values are overlaid as points. The replay advantage on retention holds at every seed with no crossovers.

## 5.2 Statistical Significance

With five paired seeds we test each pairwise difference using a paired $t$-test (Table 2). Every retention comparison is significant at $p < 10^{-3}$, including the DER++-over-ER margin of $6.3 \pm 1.1$ pp that a single run could not have separated from noise. On global accuracy the ordering is likewise significant, though the DER++-over-ER margin is much smaller ($0.7 \pm 0.4$ pp, $p = 0.018$): DER++ buys most of its retention advantage back in Task-B efficiency, so the two replay methods are far closer overall than the retention column alone suggests.

We report $t$-tests rather than the Wilcoxon signed-rank test because with $n = 5$ paired observations the smallest attainable two-sided Wilcoxon $p$-value is $2/2^5 = 0.0625$; the test cannot reach $\alpha = 0.05$ regardless of effect size, and would have declared every comparison here non-significant. We note this because it is an easy trap at small seed counts.

Table 2: Paired differences in Task-A retention across five seeds (percentage points), with paired $t$-test statistics.

| Comparison | Δ Ret. | $t$ | $p$ |
|---|---|---|---|
| DER++ − ER | $+6.3 \pm 1.1$ | 12.9 | $2 \times 10^{-4}$ |
| DER++ − MAS | $+14.5 \pm 1.2$ | 26.3 | $1 \times 10^{-5}$ |
| DER++ − EWC | $+27.1 \pm 3.3$ | 18.7 | $5 \times 10^{-5}$ |
| ER − MAS | $+8.3 \pm 1.0$ | 18.6 | $5 \times 10^{-5}$ |
| ER − EWC | $+20.9 \pm 3.0$ | 15.8 | $9 \times 10^{-5}$ |
| MAS − EWC | $+12.6 \pm 2.6$ | 10.9 | $4 \times 10^{-4}$ |

## 5.3 The Cost of Selecting on One Seed

Because the configurations were chosen to maximise retention on seed 42, that seed is expected to be optimistic. Holding the selection fixed and comparing seed 42 against the mean of the four held-out seeds quantifies the effect, and it is not uniform. EWC is inflated by $+7.4$ pp (69.2 versus 61.8), MAS by $+2.6$, ER by $+2.8$, and DER++ by only $+1.1$. The method with the widest search grid and the weakest performance is flattered most.

This has a practical consequence beyond our own numbers. Reporting a CL comparison at each method's retention-selected operating point, on the same run used for selection, systematically compresses the gap between strong and weak methods: on seed 42 the DER++-over-EWC margin reads 22.1 pp, whereas on the held-out seeds it is 28.4 pp. Selection-conditioned single-run comparisons should be read as lower bounds on the spread between methods, not as unbiased estimates.

## 5.4 Where Forgetting Concentrates

Macro retention conceals substantial heterogeneity (Table 3). Email is essentially untouched under every method, and handwritten and specification stay high, while invoice and presentation absorb most of the damage. Under EWC, invoice retains $21.8 \pm 6.0\%$ and presentation $25.1 \pm 9.9\%$ while email retains $99.6 \pm 0.6\%$ — a single macro number of 63.3% describes neither. The methods differ mainly in how far they rescue these two classes: DER++ lifts invoice to $87.3 \pm 1.6\%$ and presentation to $72.7 \pm 4.3\%$.

Per-class variance is also concentrated in the same place. EWC's specification retention varies

by $\pm 12.6$ pp across seeds, an order of magnitude more than its email retention. A practitioner reading only the macro mean would neither know which document types are at risk nor how reproducible that risk is. The retained artifacts contain per-class accuracies but not per-image predictions, so we cannot determine whether visual similarity between invoices, presentations and the Task-B classes drives the pattern; that would require a confusion-level analysis we did not archive.

Table 3: Task-A per-class retention (%), mean $\pm$ sd over five seeds. Two classes absorb most of the loss under every method.

| Method | Email | Hand. | Spec. | Invoice | Pres. |
|---|---|---|---|---|---|
| EWC | $99.6 \pm 0.6$ | $91.6 \pm 2.8$ | $78.4 \pm 12.6$ | $21.8 \pm 6.0$ | $25.1 \pm 9.9$ |
| MAS | $99.4 \pm 0.7$ | $82.9 \pm 3.4$ | $88.4 \pm 3.5$ | $63.3 \pm 4.1$ | $45.3 \pm 4.7$ |
| ER | $99.7 \pm 1.0$ | $91.6 \pm 2.7$ | $88.7 \pm 4.0$ | $77.3 \pm 3.3$ | $63.4 \pm 3.4$ |
| DER++ | $100.7 \pm 0.7$ | $95.8 \pm 2.2$ | $95.5 \pm 2.5$ | $87.3 \pm 1.6$ | $72.7 \pm 4.3$ |

### 5.5 Buffer Size and Label Budget

Both sweeps below were run on seed 42 only and are reported as trends, without the uncertainty attached to Table 1.

Retention improves monotonically with buffer size, DER++ ahead of ER at every size (Table 4). DER++ with only 200 buffered samples (40 per class, 9.6 MiB) already reaches 81.9% retention, exceeding the best regularisation configuration at under a quarter of the 42.2 MiB those regularisers spend on importance matrices. At small buffers replay is thus both stronger and cheaper; at 1,000 samples (47.9 MiB) it remains stronger but no longer cheaper, trading data storage for parameter statistics. The efficiency column shows the cost: DER++'s Task-B recovery declines from 96.7% to 94.9% as the buffer grows, so a larger buffer buys stability by conceding plasticity.

Table 4: Replay buffer sweep (seed 42). Retention rises with buffer size while DER++'s Task-B efficiency falls.

| | ER | | DER++ | |
|---|---|---|---|---|
| Buffer | Ret. | Eff. | Ret. | Eff. |
| 200 | 72.6 | 97.0 | 81.9 | 96.7 |
| 500 | 80.2 | 97.3 | 88.4 | 96.1 |
| 1,000 | 86.4 | 96.9 | 91.3 | 94.9 |

Label scarcity tells a mirror-image story (Figure 3). Re-running Phase 2 across labelled Task-B budgets from 200 to 8,000 samples, replay dominates retention at every budget, and at 200 labels all four methods retain at least 97.4% of Task-A accuracy. That apparent success is misleading: Task-B recovery at that budget spans only 49.0–72.8%, so high retention under extreme scarcity indicates limited adaptation rather than a good balance. ER is the most label-efficient new-task learner throughout, whereas DER++ is the most conservative, its logit matching over-regularising when data is scarce. Increasing the budget monotonically raises recovery and global accuracy while lowering retention, for every method.

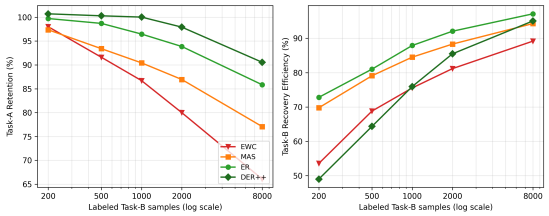


Figure 3: Task-A retention (left) and Task-B efficiency (right) versus labelled Task-B budget (200–8,000, log scale), seed 42. Near-total retention at the smallest budget reflects limited adaptation, not a favourable trade-off.

### 5.6 The Non-Distilled Control Cannot Isolate Distillation

The submitted version of this work compared the distilled student against a same-size non-distilled ViT-Tiny (ImageNet-initialised `WinKawaks/vit-tiny-patch16-224`) and read the resulting gaps as a distillation effect. That reading does not hold, and we correct it here. The non-distilled model begins Phase 1 at 6.6% accuracy on the 16 classes — essentially chance — against 90.0% for the distilled student. An 83.4 pp gap in starting accuracy confounds optimisation, representation quality and calibration simultaneously, so the comparison cannot attribute anything to distillation specifically.

What the control does show is narrower but still useful: the family ordering reproduces under a drastically weaker initialisation, with DER++ and ER retaining 82.5% and 72.7% against 51.5% and 32.7% for EWC and MAS (Table 5). The replay advantage is therefore not an artifact of the distilled starting point. Establishing a distillation effect would require a non-distilled ViT-Tiny trained to comparable 16-class accuracy before Phase 1; we did not run that control, and we make no distillation claim without it.

Table 5: Distilled versus non-distilled ViT-Tiny of identical size, single seed. The 6.6% versus 90.0% starting-accuracy gap makes this a diagnostic, not an accuracy-matched control.

| | Task-A Ret. (%) | | Global Acc. (%) | |
|---|---|---|---|---|
| Method | Distil. | Raw | Distil. | Raw |
| Start | — | — | 90.0 | 6.6 |
| No-CL | 26.5 | 8.4 | 68.1 | 55.4 |
| EWC | 69.2 | 51.5 | 75.7 | 49.2 |
| MAS | 78.0 | 32.7 | 81.6 | 56.0 |
| ER | 86.4 | 72.7 | 85.5 | 73.6 |
| DER++ | 91.3 | 82.5 | 85.8 | 72.6 |

### 5.7 Deployment Cost

Table 6 reports Phase-2 cost on a single NVIDIA T4, averaged over the four held-out seeds, with GPU energy read from the on-board NVML counter. Two points matter for practitioners.

First, once method *preparation* is counted, the time advantage of regularisation largely disappears. Estimating a Fisher or importance matrix over $\sim$9,500 Task-A samples costs 7.3–7.5 minutes, whereas constructing a replay buffer costs 1.4–1.5. Comparing training loops alone puts EWC 9.1 minutes ahead of ER; comparing total adaptation time narrows this to 3.3 minutes. Prior reports that time only the optimisation loop overstate how much cheaper regularisation is.

Second, the absolute costs are small. Every method completes an adaptation round in under 37 minutes, below 600 MiB of GPU memory, and under 26 Wh — roughly the energy of running a laptop for two hours. Distillation is performed once; each subsequent adaptation is cheap enough for institutions where annotation arrives incrementally and compute budgets are fixed.

For serving rather than adaptation, the student runs at 6.3 ms per image (batch 1) on the T4. Since the motivating deployments may lack a GPU entirely, we also measured a controlled single-thread CPU benchmark on an Intel Core i7-13700H: median model-forward latency is 61.6 ms (mean 64.9; 10th–90th percentile 45.5–89.6) over 200 timed passes after 30 warm-ups. This is one machine and one timing session, and it excludes data loading and application logic, but it establishes that CPU-only serving of the 21.2 MiB student is feasible at roughly 16 pages per second per core.

Table 6: Phase-2 deployment cost on a single NVIDIA T4, mean over four held-out seeds. Prep = method preparation (Fisher/importance estimation or buffer construction); Train = optimisation loop; $S_{\text{aux}}$ = persistent auxiliary storage; Energy from NVML.

| Method | Prep (min) | Train (min) | GPU (MiB) | $S_{\text{aux}}$ (MiB) | Energy (Wh) |
|---|---|---|---|---|---|
| Naive FT (ref.) | 0.0 | 18.7 | 358 | 0 | 13.6 |
| EWC ($\lambda$=50k) | 7.3 | 26.1 | 421 | 42.2 | 18.0 |
| MAS ($\lambda$=5.0) | 7.5 | 26.9 | 421 | 42.2 | 18.5 |
| ER (buffer=1k) | 1.4 | 35.2 | 591 | 47.9 | 25.8 |
| DER++ (buffer=1k) | 1.5 | 33.3 | 586 | 47.9 | 25.6 |

## 6 Discussion

The consistent advantage of replay over regularisation — 27 pp between DER++ and EWC, significant at every pairwise step — is consistent with how each family uses limited capacity. Regularisation penalises deviation from $\theta_{\text{naive}}$ but supplies no direct signal for the target classes during Task-B learning; the model must satisfy the Task-B loss and the penalty simultaneously. Replay instead includes Task-A samples in every batch, and DER++'s logit matching additionally preserves the output distribution. We emphasise that this is an interpretation of the observed behaviour, not an isolated mechanism: with capacity held fixed at ViT-Tiny throughout, we cannot show that compactness is what makes the indirect mechanism fail. A Tiny/Small/Base sweep would be required, and is the single most informative extension of this work.

The 12.6 pp advantage of MAS over EWC, despite their shared quadratic penalty, is notable and now significant. EWC's Fisher measures loss curvature; MAS measures sensitivity of the output norm. For a specialist confident on only 5 classes, output-based importance appears to identify the parameters essential to those predictions better than loss-curvature signals do.

For practitioners deploying distilled document classifiers that must adapt over time, our results support four guidelines. First, prefer replay: DER++ retains 90% of prior knowledge at an auxiliary storage cost ($\sim$48 MiB) comparable to what regularisation already requires, and the ordering is stable across seeds. Second, choose between the replay variants by what you need: DER++ for stability, ER if Task-B recovery matters more, since their global-accuracy difference is under a point. Third, if a replay buffer is infeasible — and page

retention may be prohibited by governance rather than by storage — MAS is clearly preferable to EWC. Fourth, do not read a macro retention number as uniform protection: budget for the two or three document types that will absorb most of the loss, and validate on those specifically.

## 7 Limitations

Several boundaries qualify these findings. The evaluation uses a single benchmark subset, a single student architecture, and a single fixed task partition and order; CL results are known to be sensitive to task composition, and multiple partitions would be needed to show the ordering is not specific to this split. The subset's sampling procedure is undocumented, so absolute accuracies should not be compared with published full-split RVL-CDIP results. Although the headline comparison now rests on five seeds with paired tests, the buffer, label-budget and non-distilled analyses remain single-seed and are reported as trends. Hyperparameters were selected on seed 42 and frozen rather than re-tuned per seed, which is what makes seeds 43–46 valid held-out evaluations but also means each method is reported at a configuration tuned on one draw; Section 5.3 quantifies the resulting optimism rather than removing it. The capacity argument that motivates the study is untested, and the non-distilled control cannot support a distillation claim. Finally, the CPU measurement is one machine and one session, and no genuine low-resource-language document collection was evaluated: our claim of transferability rests on document classification being largely layout-driven, which is an argument, not evidence.

## 8 Conclusion

We studied the interaction between knowledge distillation and continual learning for document classification. Distilling a DiT-Large teacher (307 M) into a ViT-Tiny student (5.5 M, 21.2 MiB, 90.0% accuracy) and evaluating four CL methods across five seeds, we found that replay consistently beats regularisation — DER++ $90.4 \pm 1.2\%$ retention against $75.9 \pm 2.1\%$ for the best regulariser — with every pairwise ordering significant and stable across seeds. Measuring the cost of single-seed selection showed it inflates retention by up to 7.4 pp and flatters the weakest method most. Per-class analysis showed that macro retention hides concentrated damage in two of five target classes. Counting method preparation reduced regularisation's apparent speed advantage from 9.1 to 3.3 minutes per round, with every method finishing under 37 minutes and 26 Wh on one commodity GPU.

Future work should run the capacity sweep the compactness argument requires, add an accuracy-matched non-distilled control, evaluate multiple class partitions and orders, and validate on a genuine low-resource-language document collection. Code, the distilled model, the benchmarking harness, and all seed-level result artifacts are available at `https://github.com/Mehdi-Soufiane/distil-and-evolve`.

## References


Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154. Springer.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. 2020. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems*, volume 33, pages 15920–15930.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Tal Hassner, and Louis-Philippe Morency. 2019. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. In *Advances in Neural Information Processing Systems*, volume 32.

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

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *Proceedings of the 2015 International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for document AI with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091. ACM.

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

Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. 2022. DiT: Self-supervised pre-training for document image transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3530–3539. ACM.

Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947.

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

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. 2017. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2001–2010. IEEE.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. 2023. CODA-Prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11909–11919.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. 2022. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2021. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 2579–2591. ACL.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200. ACM.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. *Proceedings of the 34th International Conference on Machine Learning*, 70:3987–3995.