MITIGATING CATASTROPHIC FORGETTING IN LLM CONTINUAL LEARNING THROUGH ATTRACTOR-INFORMED LATENT TRANSITION ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual learning enables large language models to acquire new knowledge while preserving the computational structures that support their existing capabilities. In replay-free continual learning, however, original pre-training data are unavailable, and model updates may disrupt these structures and cause catastrophic forgetting. Existing methods mainly preserve model parameters, output distributions, or isolated hidden states, while paying less attention to cross-layer transformations. We propose Latent Transition Alignment (LTA), which uses cross-layer latent transitions as a preservation target. In selected intermediate Transformer windows, our latent-dynamics analysis identifies coherent transition fields and locally attracting fixed points. We then introduce an offline dual-anchor surrogate that connects observed transitions to the residual field of a latent self-map. Under exact reconstruction, the observed latent transition equals the residual-field vector at its start state. Under approximate reconstruction, their discrepancy is bounded. Online, LTA aligns transition directions and applies one-sided magnitude control without online field fitting or fixed-point iteration. In single-domain learning, LTA outperforms all evaluated baselines in retention while matching SFT in new-domain learning; relative to Output KL, it reduces accumulated first-domain forgetting by 8.20 percentage points and improves final QASC accuracy by 7.30 percentage points in multi-domain continual learning; under fixed-protocol transfer to Qwen2.5-3B, it requires no scale-specific retuning and reduces forgetting by 63.1% while improving QASC accuracy by 3.33 percentage points.

1 INTRODUCTION

Large language models must acquire new knowledge while retaining existing capabilities. We study replay-free continual learning, where adaptation proceeds without access to original pre-training examples. Existing preservation objectives constrain parameters, predictions, or hidden representations (Ke et al., 2023; Jiang et al., 2025). However, Transformer computation repeatedly transforms representations across depth. Preserving an output or an isolated hidden state does not fully specify these transformations.

Intermediate and cross-layer distillation show that feature evolution contains transferable knowledge (Sun et al., 2019; Gong et al., 2025). We examine cross-layer latent transitions in selected intermediate Transformer windows. Following the latent-dynamics perspective of Fumero et al. (2026), our analysis identifies coherent transition fields and locally attracting fixed points. Constraining transition-field structure reduces forgetting while improving new-domain learning. These findings motivate preserving cross-layer relationships without fixing representation locations. Matching the complete transition, however, constrains both direction and magnitude and can restrict useful adaptation.

We introduce Latent Transition Alignment (LTA). An offline dual-anchor surrogate connects observed transitions to the residual field of a latent self-map. We establish an transition–field identi ty under exact reconstruction and bound the discrepancy under approximate reconstruction.Onlin e, frozen encoders extract latent transitions from the reference and trainable models.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

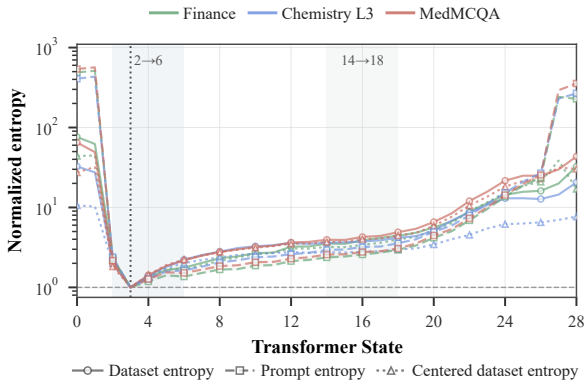

Figure 1: **Layer-wise spectral entropy in the frozen Qwen2.5-1.5B model.** Curves for Finance, Chemistry L3, and MedMCQA are normalized to their values at state 3. Shaded regions mark the selected windows $2 \to 6$ and $14 \to 18$.

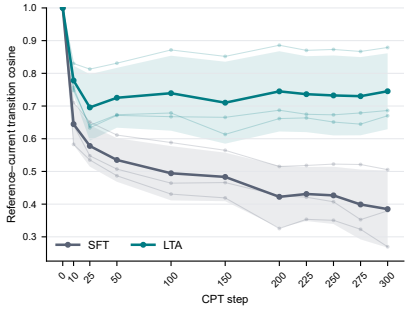

Figure 2: **Transition alignment during NanoChat adaptation to TinySci-Fiction.** Thick lines show three-seed means, faint lines individual runs, and shading sample standard deviation.

LTA aligns their directions and penalizes only excessive magnitude growth, allowing common endpoint shifts and smaller displacements. It requires no online field fitting or fixed-point iteration and adds no inference-time components.

2 Related Work

Continual learning. Existing approaches retain prior knowledge through output preservation (Li & Hoiem, 2016), importance-based soft masking (Ke et al., 2023), or orthogonal low-rank subspaces (Wang et al., 2023). Function-vector-guided training combines activation consistency with an intervention-guided KL objective (Jiang et al., 2025). Pathway protection and RECALL use graph-matching-based fusion and representation-guided hierarchical merging, respectively (Chen et al., 2025; Wang et al., 2025).

Intermediate and cross-layer distillation. FitNets and Patient Knowledge Distillation transfer intermediate representations (Romero et al., 2015; Sun et al., 2019). Instance Relationship Graph distillation captures inter-instance relationships and cross-layer transformations (Liu et al., 2019). Feature Dynamics Distillation (FDD) aligns feature trajectories and uses cosine matching of layer-wise changes after projection through language-model heads (Gong et al., 2025). MTA aligns representations and relational structure at depth-dependent semantic granularities (Chi et al., 2026). LTA applies cross-layer alignment to replay-free continual learning through a dual-anchor transition–field connection, directional alignment, and one-sided magnitude control.

Neural networks as dynamical systems. Neural ODEs interpret depth through continuous dynamics (Chen et al., 2018). Deep equilibrium models compute fixed-point representations, while recurrent-network studies use fixed-point analysis to reveal computational mechanisms (Bai et al., 2019; Maheswaranathan et al., 2019). Fumero et al. (2026) study latent vector fields induced by encoder–decoder iteration.

3 Motivation

3.1 From Pointwise Representations to Intermediate Cross-Layer Transformations

Two models can produce nearby representations while implementing different transformations around them. Conversely, representations can move while preserving useful cross-layer relationships. Building on evidence that intermediate representations and feature dynamics carry transfer-

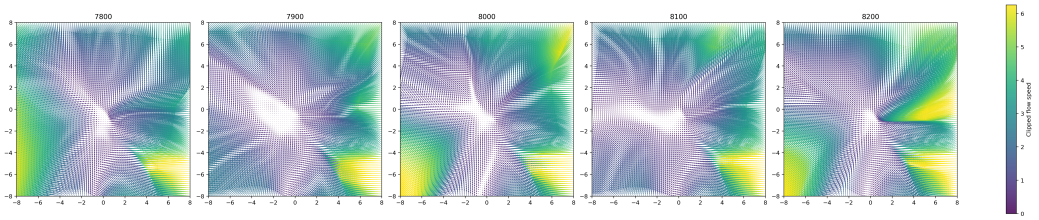

Figure 3: **NanoChat latent transition fields across pre-training checkpoints.** Panels show the layer-$4 \to 5$ fields at steps 7,800, 7,900, 8,000, 8,100, and 8,200. Arrows show direction and colors show clipped field magnitudes. Adjacent-field cosines range from 0.960 to 0.986 over common support.

able knowledge (Sun et al., 2019; Yim et al., 2017; Gong et al., 2025), we study forgetting as a change in how representations are transformed, rather than only where they are located.

Layer-wise analyses show that representation geometry and computational contributions vary systematically with depth (Razzhigaev et al., 2024; Skean et al., 2025; Csordás et al., 2025). In the frozen Qwen2.5-1.5B model, Finance, Chemistry L3, and MedMCQA inputs (Pal et al., 2022) exhibit a shared compression–re-expansion profile under three measurements: dataset spectral entropy, mean per-prompt token spectral entropy, and centered dataset spectral entropy (Figure 1). The recurring trough across these domains and measurements identifies a stable depth organization: representations first become concentrated and subsequently re-expand. This structure motivates examining windows spanning feature aggregation and subsequent reorganization, where intermediate representations are consolidated and passed to later computation.

Let $H_\ell^m(x) \in \mathbb{R}^{T \times d_h}$ denote the hidden-state matrix of a sequence x at depth ℓ in model version m. Its token state at position u is $h_{\ell,u}^m(x) = H_\ell^m(x)[u,:]$. The superscript m identifies a model checkpoint or its role as the initial base, stage reference, or trainable model. For window $r = (a_r, b_r)$, the Transformer subnetwork satisfies

$$H_{b_r}^m(x) = \Phi_{a_r \to b_r}^m \left(H_{a_r}^m(x) \right). \tag{1}$$

We next ask what stable organization these cross-layer transformations contain and why preserving it can mitigate forgetting.

3.2 Latent Dynamics Reveal Stable Cross-Layer Organization

High-dimensional hidden states combine token-level variation, contextual detail, and the structure supporting cross-layer computation.

To examine their collective transition patterns, we train a shared low-dimensional encoder $\mathcal{E}_r$ and decoder $\mathcal{D}_r$ on frozen-model states. Both endpoints use the same coordinates, $\widetilde{z}_{\ell,u}^m = \mathcal{E}_r(h_{\ell,u}^m(x))$, $\ell \in \{a_r, b_r\}$, and are reconstructed separately: $\mathcal{D}_r(\widetilde{z}_{a_r,u}^m) \approx h_{a_r,u}^m(x)$ and $\mathcal{D}_r(\widetilde{z}_{b_r,u}^m) \approx h_{b_r,u}^m(x)$.

For a frozen checkpoint τ, we fit a residual field to its observed transitions: $\widetilde{\Delta z}_{r,u}^{\,\tau} = \widetilde{z}_{b_r,u}^{\tau} - \widetilde{z}_{a_r,u}^{\tau}$, $\widehat{V}_{r,\tau}(\widetilde{z}_{a_r,u}^{\tau}) \approx \widetilde{\Delta z}_{r,u}^{\,\tau}$, and $\widehat{g}_{r,\tau}(z) = z + \widehat{V}_{r,\tau}(z)$.

Following the latent-dynamics perspective developed by Fumero et al. (2026), we examine the self-map through its fixed points and their local stability. Iterating $\widehat{g}_{r,\tau}$ reveals how nearby states converge toward stable patterns of the fitted cross-layer dynamics.

We visualize this organization in an eight-layer, approximately 100M-parameter NanoChat model with hidden width 512 (Karpathy, 2025), pretrained on TinyStories (Eldan & Li, 2023). A two-dimensional reconstruction space represents the fourth-to-fifth-layer transition. A single frozen diagnostic encoder defines the latent coordinates shared by all analyzed checkpoints. Holding the observation samples fixed, we refit only the checkpoint-specific residual heads. Adjacent fields have cosine similarities of 0.960–0.986 over common support (Figure 3), revealing stable collective transition patterns as pre-training continues.

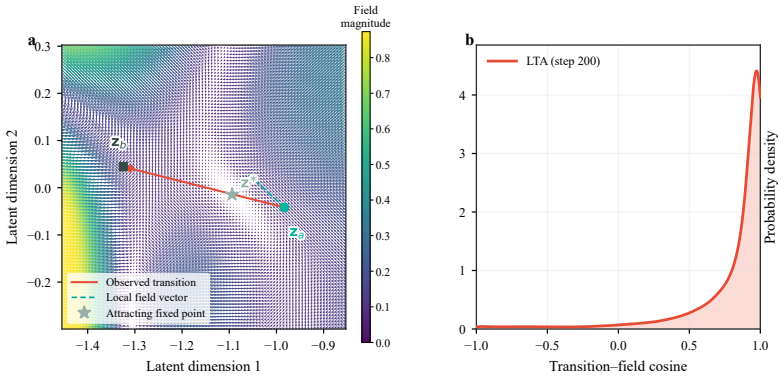

Figure 4: **Observed transitions and local latent dynamics in NanoChat.** (a) A held-out transition, its corresponding fitted residual-field vector, and a nearby attracting fixed point. (b) The distribution of transition–field cosine over held-out tokens for LTA at adaptation step 200, with a median of 0.940.

Independent fixed-point searches, local Jacobian analysis, and perturbation-return tests identify locally attracting fixed points. The observed transitions also agree with the fitted field on held-out tokens, with a median cosine similarity of 0.940 (Figure 4). Together, these results connect the actual one-step cross-layer computation to the stable local dynamics observed in the reconstruction space.

To test the functional relevance of these relationships, we adapt the model to TinySciFiction. We construct this structure-matched domain from TinyStories using DeepSeek V4 to replace fairy-tale vocabulary and entities with simple science-fiction concepts in batches. We compare standard next-token SFT with a paired intervention that aligns transition directions and limits excessive transition magnitude. Both use the same pretrained initialization, batch order, and evaluation checkpoints. At step 300, the three-seed mean transition cosine increases from 0.385 to 0.745 (Figure 2). ForgetNLL decreases from 1.2447 to 0.8745, a 29.75% reduction, while new-domain NLL gain increases from 1.0036 to 1.2610. Preserving cross-layer relationships thus reduces forgetting without suppressing new-domain learning.

The practical question is how to preserve these relationships during continual learning without repeatedly fitting a complete vector field and recovering its attracting structure.

4 Method

Latent Transition Alignment (LTA) replaces repeated field fitting, latent rollouts, and stability analysis with constraints on the model's observed cross-layer transitions. Offline, a dual-anchor surrogate closes the window transformation into a self-map on the window-input space and connects its residual field to an actual transition. Online, frozen encoders extract matched reference and trainable transitions for direction alignment and one-sided magnitude control (Figure 5).

4.1 Offline Dual-Anchor Construction and the Transition–Field Connection

The reconstruction modules in Section 3.2 recover each endpoint for dynamical observation. For LTA, we instead train a window-specific encoder E_r and start-state decoder D_{a_r}. The encoder maps both endpoints into shared latent coordinates, while the decoder associates both codes with the same window input. On endpoint pairs extracted from the initial frozen base model M^{base}, they are jointly trained with

$$\mathcal{L}^{(r)}_{\text{surr}} = \mathbb{E}_{(x,u)}\Big[\big\|D_{a_r}(E_r(h^{\text{base}}_{a_r,u}(x))) - h^{\text{base}}_{a_r,u}(x)\big\|_2^2 + \big\|D_{a_r}(E_r(h^{\text{base}}_{b_r,u}(x))) - h^{\text{base}}_{a_r,u}(x)\big\|_2^2\Big]. \tag{2}$$

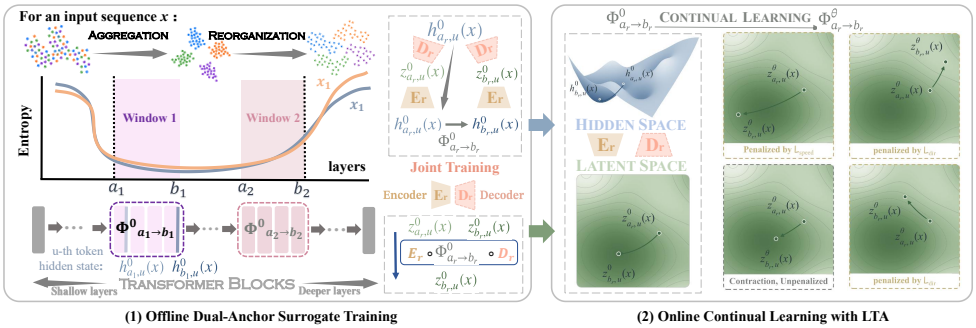

Figure 5: **Overview of LTA. Left:** Frozen-base entropy guides the selection of Transformer windows spanning feature aggregation and reorganization. A jointly trained dual-anchor encoder–decoder closes each window transformation into a latent self-map. **Right:** LTA aligns reference and current transition directions and penalizes excessive magnitude growth. The one-sided magnitude term leaves contractions unpenalized.

For token position u, let c_u collect the layer-a_r states of all positions other than u, together with the causal mask and positional metadata. We write $\Phi^m_{a_r\to b_r}(h; c_u)$ for the window output at that position after inserting the candidate state h into this context. The latent self-map and its corresponding window-input self-map are

$$\begin{aligned} g^m_{r,c_u}(z) &= E_r\big(\Phi^m_{a_r\to b_r}(D_{a_r}(z); c_u)\big)\,, \\ \mathcal{R}^m_{r,c_u}(h) &= D_{a_r}\big(E_r(\Phi^m_{a_r\to b_r}(h; c_u))\big)\,. \end{aligned} \tag{3}$$

Their fixed points correspond through D_{a_r} and $E_r \circ \Phi^m_{a_r\to b_r}$. The appendix derives the trajectory and local-stability relationships.

Let $z^{\text{base}}_{a_r,u} = E_r(h^{\text{base}}_{a_r,u}(x))$ and $z^{\text{base}}_{b_r,u} = E_r(h^{\text{base}}_{b_r,u}(x))$. Under exact dual-anchor reconstruction,

$$g^{\text{base}}_{r,c_u}(z^{\text{base}}_{a_r,u}) = z^{\text{base}}_{b_r,u}, \qquad g^{\text{base}}_{r,c_u}(z^{\text{base}}_{b_r,u}) = z^{\text{base}}_{b_r,u}. \tag{4}$$

Defining $V^{\text{base}}_{r,c_u}(z) = g^{\text{base}}_{r,c_u}(z) - z$ gives the key connection:

$$\boxed{V^{\text{base}}_{r,c_u}(z^{\text{base}}_{a_r,u}) = z^{\text{base}}_{b_r,u} - z^{\text{base}}_{a_r,u} = \Delta z^{\text{base}}_{r,u}.} \tag{5}$$

With approximate reconstruction, the discrepancy is bounded by the start-state reconstruction error and the local sensitivity of the encoded window, as detailed in the appendix.

After offline training, we freeze both modules and their coordinate normalization. Only the encoder is used online. The decoder remains an analysis component. The appendix gives the architectures and detailed training protocol.

4.2 Online Latent Transition Alignment

At continual-learning stage t, a trainable model M^θ_t is initialized from the frozen reference M^{ref}_t. The first-stage reference is $M^{\text{ref}}_1 = M^{\text{base}}$. Both models process the same current-domain sequence and causal mask, without replaying original pre-training examples. Their matched transitions are

$$\begin{aligned} \Delta z^{\text{ref},t}_{r,u} &= \text{sg}\Big[E_r(h^{\text{ref},t}_{b_r,u}(x)) - E_r(h^{\text{ref},t}_{a_r,u}(x))\Big]\,, \\ \Delta z^{\theta,t}_{r,u} &= E_r(h^{\theta,t}_{b_r,u}(x)) - E_r(h^{\theta,t}_{a_r,u}(x)), \end{aligned} \tag{6}$$

where sg denotes stop-gradient. Shared inputs and encoders make these paired differences a measure of changes in the model's cross-layer relationships.

LTA separates direction from magnitude. Direction specifies the orientation of representation transport, while magnitude measures the size of a cross-layer displacement. Full-transition matching penalizes both increases and decreases in magnitude. LTA instead preserves direction and constrains only excessive growth, leaving room to adjust transition strength when learning new data. Both transition-based objectives permit common endpoint shifts. LTA additionally treats magnitude changes asymmetrically.

We suppress the minibatch index, so the sums over u include all token positions in the current minibatch. Let $\mu_u \in \{0, 1\}$ mark valid causal-token positions and $N = \sum_u \mu_u$. We use the stabilized cosine and dimension-normalized transition magnitude

$$\begin{aligned} \cos_\epsilon(a, b) &= \frac{\langle a, b\rangle}{\max(\|a\|_2, \epsilon)\max(\|b\|_2, \epsilon)}, \\ v(\Delta z) &= \frac{\sqrt{\|\Delta z\|_2^2 + \epsilon}}{\sqrt{d_z}}. \end{aligned} \tag{7}$$

A reference-magnitude weight emphasizes states with smaller reference displacements. Its values depend only on the frozen reference branch. Together with the direction and magnitude terms for window r, these quantities are

$$w_{r,u}^{\mathrm{ref},t} = \exp\left[-\beta v(\Delta z_{r,u}^{\mathrm{ref},t})^2\right], \tag{8}$$

$$\mathcal{L}_{\mathrm{dir}}^{(r)} = \frac{\lambda_{\mathrm{dir}}}{N}\sum_u \mu_u w_{r,u}^{\mathrm{ref},t}\left[1 - \cos_\epsilon\left(\Delta z_{r,u}^{\theta,t}, \Delta z_{r,u}^{\mathrm{ref},t}\right)\right], \tag{9}$$

$$\mathcal{L}_{\mathrm{mag}}^{(r)} = \frac{\lambda_{\mathrm{mag},r}}{N}\sum_u \mu_u w_{r,u}^{\mathrm{ref},t}\left[v(\Delta z_{r,u}^{\theta,t}) - \gamma v(\Delta z_{r,u}^{\mathrm{ref},t})\right]_+^2, \tag{10}$$

where $[s]_+ = \max(s, 0)$, $\epsilon > 0$, $\beta \geq 0$, and $\gamma > 1$. The coefficients λ_{dir} and $\lambda_{\mathrm{mag},r}$ control the relative strengths of the direction and magnitude penalties.

For R selected windows $\mathcal{W} = \{(a_r, b_r)\}_{r=1}^R$, the training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{LM}} + \alpha\sum_{r=1}^R \omega_r\left(\mathcal{L}_{\mathrm{dir}}^{(r)} + \mathcal{L}_{\mathrm{mag}}^{(r)}\right), \tag{11}$$

where $\mathcal{L}_{\mathrm{LM}}$ is next-token cross-entropy, ω_r weights windows, and α sets the overall regularization strength. Encoder parameters remain frozen, while the trainable branch retains gradients through E_r to its hidden-state inputs. Online updates use only the two model forwards, frozen encoders, and observed endpoints.

For sequential learning, the model completing stage $t-1$ is copied and frozen as the rolling reference M_t^{ref} for $t > 1$. The encoders retain their initial offline versions. After adaptation, only the updated language model is retained, with no additional inference-time component.

5 Experiments

We evaluate Latent Transition Alignment (LTA) by asking four questions. First, are cross-layer transitions a more effective preservation target than outputs or pointwise hidden states? Second, do selective constraints on transition direction and one-sided magnitude outperform full-transition matching? Third, does the resulting advantage persist across different task orders? Fourth, does this advantage transfer to Qwen2.5-3B? Throughout the experiments, we report target-domain learning and reference-behavior retention separately to avoid mistaking inhibited learning for reduced forgetting.

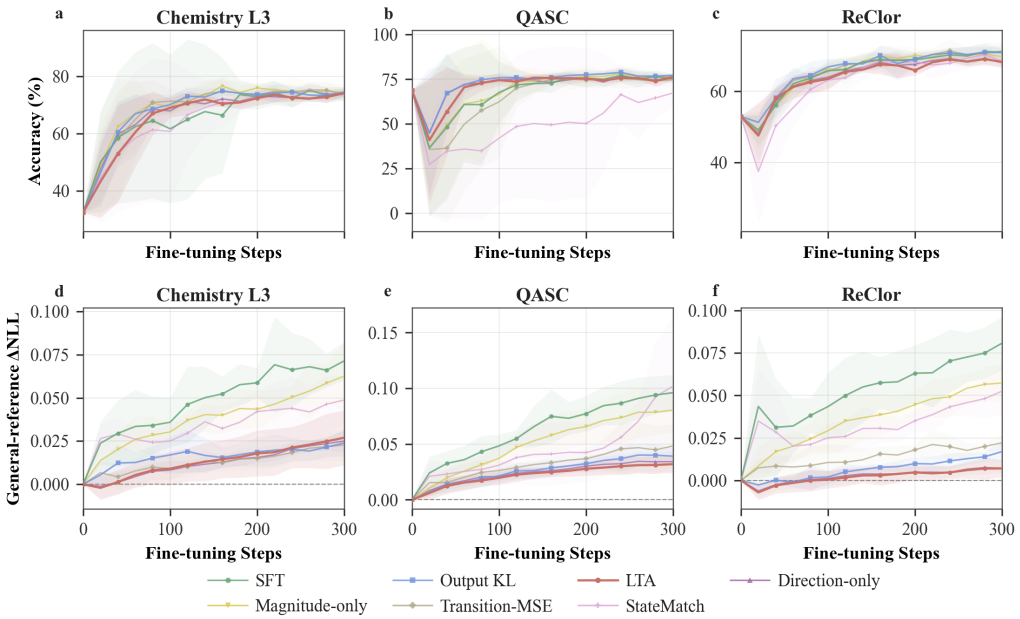

Figure 6: **Single-domain learning and retention on Qwen2.5-1.5B.** Columns show Chemistry L3, QASC, and ReClor. The top row reports development-set accuracy, and the bottom row reports reference ΔNLL relative to the frozen base model. Curves show means over five paired seeds, with shaded 95% Student's t confidence intervals.

5.1 Experimental Setup

Models, datasets, and protocol. Our main experiments used Qwen2.5-1.5B (Qwen Team, 2024) on Chemistry L3, QASC, and ReClor (Feng et al., 2024; Khot et al., 2020; Yu et al., 2020). We evaluated three sequential orders: QASC $\rightarrow$ Chemistry $\rightarrow$ ReClor, ReClor $\rightarrow$ Chemistry $\rightarrow$ QASC, and ReClor $\rightarrow$ QASC $\rightarrow$ Chemistry. The scale experiment used Qwen2.5-3B on QASC. All formal Qwen comparisons used five paired seeds. Methods shared the same initialization, batches, token masks, and evaluation examples within each seed. Each domain was trained for 300 steps without replaying original pre-training examples during continual adaptation. We used AdamW with a learning rate of 4×10^{-5}, an effective batch size of 32, a sequence length of 512, and gradient clipping at 30. Forward and backward passes used BF16. Structural losses and metrics were accumulated in FP32.

Window selection and surrogate training. Before continual learning, we selected intermediate Transformer windows using layer-wise representation entropy from the frozen base model. Qwen2.5-1.5B used windows $2 \rightarrow 6$ and $14 \rightarrow 18$. Qwen2.5-3B used windows $3 \rightarrow 8$ and $18 \rightarrow 23$. Each window r used a window-specific encoder E_r shared by both endpoints and a start-state decoder D_{a_r}. Both were GELU MLPs. The encoder mapped $d_h \rightarrow 256 \rightarrow 128$, and the decoder mapped $128 \rightarrow 256 \rightarrow d_h$. We trained them using the objective in Eq. (2). After 3,000 training steps, we froze the encoders and shared them across all structural methods and seeds. All window and surrogate configurations were fixed before inspecting formal continual-learning adaptation or retention outcomes.

Comparisons and metrics. We compared SFT, Output KL (Li & Hoiem, 2016), Hidden-State Matching (StateMatch) (Romero et al., 2015; Sun et al., 2019), Full-Transition Matching (Transition-MSE), a matched structural baseline motivated by feature-dynamics alignment (Gong et al., 2025), and LTA. The single-domain experiments also included the Direction-only and Magnitude-only ablations. All methods used the same reference model, paired seeds, and batches. Structural methods additionally shared the same encoders and Transformer windows. Target-domain learning was measured by development-set accuracy. For single-domain learning, let s denote the optimization step and $M^{\theta}_{1,s}$ the trainable model after s updates. Single-domain retention was measured on a fixed reference set relative to the frozen base model: $\Delta\mathrm{NLL}_{\mathrm{ref}}(s) =$

$\mathrm{NLL}_{\mathrm{ref}}(M^{\theta}_{1,s}) - \mathrm{NLL}_{\mathrm{ref}}(M^{\mathrm{base}})$. Here, the subscript ref denotes the fixed evaluation set, and lower values indicate stronger retention. For an order o, let $A^{(o)}_{1,t}$ denote accuracy on its first domain after completing stage t. We defined accumulated forgetting as $F^{(o)}_1 = A^{(o)}_{1,1} - A^{(o)}_{1,3}$. Before each stage, we copied its input model and froze that copy as the rolling reference. For Qwen2.5-3B, we operationalized general-capability drift as aggregate ΔNLL averaged over C4 (Raffel et al., 2020), WikiText (Merity et al., 2016), and a fixed general-reference set. Performance plots for the formal Qwen experiments report means over five paired seeds, with shaded 95% Student's t confidence intervals.

5.2 Single-Domain Learning

In the single-domain experiments in Figure 6, LTA achieved the lowest final three-domain mean reference drift, at 0.0220. Relative to Output KL, Transition-MSE, and SFT, it reduced reference drift by 17.6%, 29.2%, and 73.5%, respectively. LTA simultaneously maintained competitive target-domain accuracy. On ReClor, Transition-MSE exceeded LTA by approximately three accuracy points, but incurred more than three times LTA's reference drift. Matching the complete transition vector therefore preserved reference behavior less effectively than aligning transition direction with a one-sided magnitude constraint.

The ablations further support the complete LTA objective. Direction-only was consistently the component ablation closest to full LTA, showing that transition direction carries the primary stable structure preserved by LTA. Magnitude-only produced substantially larger drift, demonstrating that magnitude control alone is insufficient. Adding one-sided magnitude control to directional alignment yielded the lowest three-domain mean drift. StateMatch was weaker overall, especially on QASC. These results show that preserving how representations evolve across depth mitigates forgetting more effectively than directly anchoring the representations themselves. Directional alignment and one-sided magnitude control also leave sufficient flexibility for learning new data. LTA can therefore mitigate forgetting without suppressing target-domain adaptation.

5.3 Sequential Learning across Three Task Orders

We next evaluated sequential domain transfer. Figure 7 reports first-domain accuracy before training and after each domain. The first post-training point measures first-task acquisition. Subsequent declines measure accumulated forgetting caused by later domain updates. Under QASC → Chemistry → ReClor, LTA retained 51.13% final QASC accuracy. This was 7.30 percentage points above Output KL, and LTA reduced accumulated first-domain forgetting by 8.20 percentage points.

We then placed ReClor first and changed the order of the two later domains. LTA reduced accumulated forgetting by 3.00 and 5.36 percentage points under ReClor → Chemistry → QASC and ReClor → QASC → Chemistry, respectively. It achieved the highest final first-domain accuracy in all three evaluated orders. LTA does not directly preserve a task-specific output distribution. Instead, it preserves cross-layer latent transition relationships between selected Transformer depths.

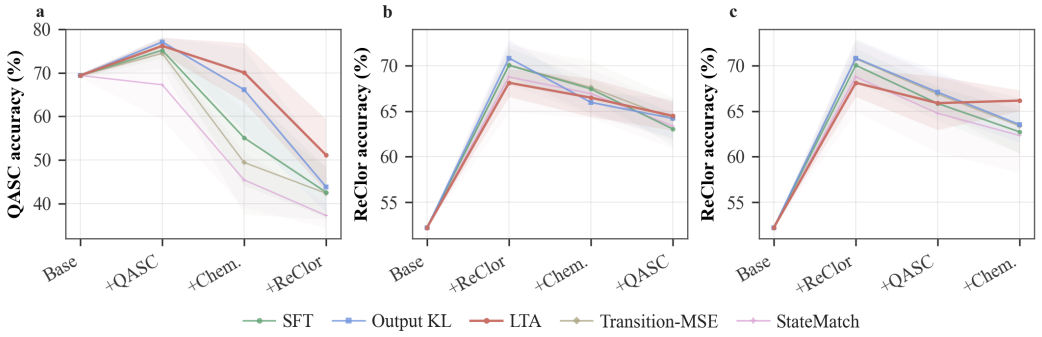

Figure 7: **First-domain retention across three continual-learning orders.** Each panel tracks first-domain accuracy before training and after each successive domain, following the order shown on the horizontal axis. Points show means over five paired seeds, with shaded 95% Student's t confidence intervals.

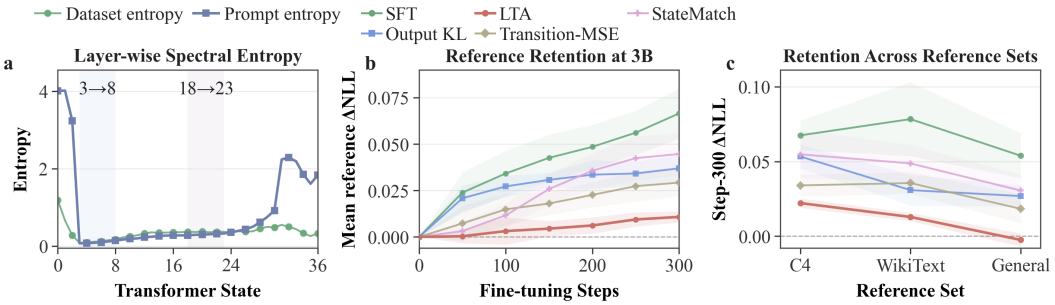

Figure 8: **Window selection and reference retention for Qwen2.5-3B on QASC.** (a) Frozen-base spectral entropy and the selected windows. (b) Mean ΔNLL over C4, WikiText, and the fixed general-reference set. (c) Per-reference-set ΔNLL at step 300. Panels (b) and (c) show means over five paired seeds, with shaded 95% Student's t confidence intervals.

The advantage was reproduced across different training orders, demonstrating that latent transition relationships provide a more transferable stability constraint than the compared baselines.

5.4 Scaling to Qwen2.5-3B

We extended LTA to Qwen2.5-3B on QASC using the same 300-step schedule and five paired seeds. The 3B experiments reuse the 1.5B optimization, surrogate-training, and LTA calibration protocols without outcome-based retuning; only the windows are selected separately from frozen-base diagnostics. Figure 8(a) shows the frozen-base layer-wise entropy diagnostic and the selected windows. Figure 8(b) reports mean ΔNLL over C4, WikiText, and the fixed general-reference set. The five-seed mean LTA curve had the lowest drift at every non-zero evaluation checkpoint. At step 300, its aggregate ΔNLL was 0.0108, which was 63.1% lower than Transition-MSE, the strongest non-LTA retention baseline.

This retention advantage did not arise from failing to learn QASC. LTA reached 77.17% final QASC accuracy, a 3.33-point gain over the base model's 73.83%. Output KL achieved the highest accuracy at 78.77%, but its mean drift across the three reference sets was 0.0370, or 3.42 times LTA's drift. LTA therefore obtained a substantial target-domain gain while reducing general-capability drift below all compared baselines.

6 Discussion

The results show that preserving cross-layer relationships provides a better retention–adaptation trade-off than fixing endpoint locations. StateMatch anchors endpoints, whereas Transition-MSE matches the complete displacement. LTA preserves direction while allowing smaller displacements. The ablations identify directional alignment as the main source of retention gains, complemented by one-sided magnitude control.

The latent-dynamics analysis identifies coherent transition fields and locally attracting fixed points. The dual-anchor construction connects observed transitions to local residual-field vectors, providing an online-computable preservation target. Retention gains persist across three task orders and transfer to Qwen2.5-3B.

7 Conclusion

We presented LTA, which uses cross-layer latent transitions as a preservation target for replay-free continual learning of large language models. Its dual-anchor construction establishes an exact transition–field connection under exact reconstruction. Directional alignment and one-sided magnitude control preserve cross-layer relationships without fixing endpoint locations. Controlled, single-domain, sequential, and fixed-protocol scale-transfer experiments show reduced forgetting with competitive target-domain learning. LTA requires no online field fitting or fixed-point iteration and adds no inference-time components.

AI USE STATEMENT

Generative AI tools, including ChatGPT and Codex, were used to assist with manuscript drafting and revision, translation, literature search, code and LaTeX preparation, and discussions of the theoretical framework, methodology, experimental design, and interpretation of results. AI-generated elements were also used in schematic illustrations. DeepSeek V4 was used to construct TinySci-Fiction from TinyStories by replacing fairy-tale vocabulary and entities with simple science-fiction concepts while retaining a similar narrative structure. The authors take responsibility for the final manuscript, including its text, mathematical claims, references, code, figures, and reported results.

REPRODUCIBILITY STATEMENT

The supplementary material provides implementations of LTA and all evaluated baselines. It includes the training and evaluation runners for the three sequential task orders and the Qwen2.5-3B experiment, together with the scripts used to reproduce Figures 6–8. Fixed experimental protocols, five-seed source tables, and automated checks for the reported experiment matrices, evaluation grids, aggregate values, and confidence intervals are also provided. Large pretrained models, derived checkpoints, and externally distributed datasets are not redistributed. Their identities and acquisition sources are specified in the paper and supplementary documentation.

REFERENCES

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Zhikang Chen, Abudukelimu Wuerkaixi, Sen Cui, Haoxuan Li, Ding Li, Jingfeng Zhang, Bo Han, Gang Niu, Houfang Liu, Yi Yang, Sifan Yang, Changshui Zhang, and Tianling Ren. Learning without isolation: Pathway protection for continual learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9377–9399. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25bt.html.

Pham Khanh Chi, Quoc Phong Dao, Thuat Nguyen, Linh Ngo Van, Trung Le, and Thanh Hong Nguyen. MTA: Multi-granular trajectory alignment for large language model distillation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32669–32684. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1507. URL https://aclanthology.org/2026.acl-long.1507/.

Róbert Csordás, Christopher D. Manning, and Christopher Potts. Do language models use their depth efficiently? In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/eae3af0f5868f0a2eceb74208966d55b-Abstract-Conference.html.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english?, 2023. URL https://arxiv.org/abs/2305.07759.

Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. SciKnowEval: Evaluating multi-level scientific knowledge of large language models, 2024. URL https://arxiv.org/abs/2406.09098.

Marco Fumero, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. Navigating the latent space dynamics of neural models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2505.22785.

Guoqiang Gong, Jiaxing Wang, Jin Xu, Deping Xiang, Zicheng Zhang, Leqi Shen, Yifeng Zhang, Junhua Shu, Zhaolong Xing, Zhen Chen, Pengzhang Liu, and Ke Zhang. Beyond logits: Aligning

feature dynamics for effective knowledge distillation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23067–23077. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1125. URL https://aclanthology.org/2025.acl-long.1125/.

Gangwei Jiang, Caigao Jiang, Zhaoyi Li, Siqiao Xue, Jun Zhou, Linqi Song, Defu Lian, and Ying Wei. Unlocking the power of function vectors for characterizing and mitigating catastrophic forgetting in continual instruction tuning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=gc8QAQfXv6.

Andrej Karpathy. nanochat: The best chatgpt that $100 can buy, 2025. URL https://github.com/karpathy/nanochat.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual pretraining of language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=m_GDIItaI3o.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. QASC: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8082–8090, 2020. doi: 10.1609/aaai.v34i05.6319.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision*, pp. 614–629. Springer, 2016. doi: 10.1007/978-3-319-46493-0_37.

Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. Knowledge distillation via instance relationship graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7096–7104, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Liu_Knowledge_Distillation_via_Instance_Relationship_Graph_CVPR_2019_paper.html.

Niru Maheswaranathan, Alex H. Williams, Matthew D. Golub, Surya Ganguli, and David Sussillo. Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016. URL https://arxiv.org/abs/1609.07843.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Qwen Team. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://www.jmlr.org/papers/v21/20-074.html.

Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. The shape of learning: Anisotropy and intrinsic dimensions in transformer-based models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 868–874. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-eacl.58. URL https://aclanthology.org/2024.findings-eacl.58/.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6550.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann Lecun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55854–55875. PMLR, 2025. URL https://proceedings.mlr.press/v267/skean25a.html.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for BERT model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 4323–4332. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1441.

Bowen Wang, Haiyuan Wan, Liwen Shi, Chen Yang, Peng He, Yue Ma, Haochen Han, Wenhao Li, Tiao Tan, Yongjian Li, Fangming Liu, Yifan Gong, and Sheng Zhang. RECALL: REpresentation-aligned catastrophic-forgetting ALLeviation via hierarchical model merging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16381–16395. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.829.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10658–10671. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.715.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4133–4141, 2017. doi: 10.1109/CVPR.2017.754. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Yim_A_Gift_From_CVPR_2017_paper.html.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. ReClor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJgJtT4tvB.

A Appendix Scope, Evidence Levels, and Notation

This appendix supplies the theoretical details, data construction and selection record, surrogate training protocol, controlled baselines, optimization contract, reproducibility map, and endpoint tables underlying the main paper. It distinguishes three kinds of evidence. *Formal paired comparisons* use identical initialization, ordered batches, masks, evaluation manifests, and five paired seeds. *Development audits* determine frozen design choices before the corresponding formal outcomes are inspected. *Mechanistic diagnostics* examine the latent construction and its local behavior; they do not by themselves establish method superiority.

Let M^{ref} be the frozen reference model at the beginning of a continual-learning stage and let $M^{\theta} \leftarrow M^{\mathrm{ref}}$ be the trainable model. Both receive the same current-domain sequence; no original pretraining example is replayed. For window $r = (a_r, b_r)$, token u, and model $m \in \{\mathrm{ref}, \theta\}$, $h^m_{a_r,u}$ and $h^m_{b_r,u}$ denote the two endpoint states. A frozen encoder $E_r : \mathbb{R}^{d_h} \to \mathbb{R}^{d_z}$ maps both endpoints into a shared latent chart. The TinyStories–TinySciFiction intervention uses three seeds $(0, 1, 2)$. Every formal Qwen comparison uses the same five paired seeds $(0, 1, 11, 271828, 141)$. Unless stated otherwise, uncertainty is a two-sided 95% Student-t confidence interval over these five observations.

B Two Latent Constructions and Their Roles

The paper uses two related but distinct latent constructions. Keeping them separate is essential because they answer different questions.

Observational reconstruction chart. For the field visualizations, an encoder $\mathcal{E}_r$ and decoder $\mathcal{D}_r$ reconstruct each endpoint in its own hidden-state space. A checkpoint-specific residual head $\widehat{V}_{r,\tau}$ is then fitted on frozen hidden-state pairs. With $\widetilde{z}_a = \mathcal{E}_r(h_a)$ and $\widetilde{z}_b = \mathcal{E}_r(h_b)$, the empirical transition

is $\widetilde{z}_b - \widetilde{z}_a$. This chart is used only to observe field organization across checkpoints and to perform local fixed-point diagnostics.

Dual-anchor LTA chart. For LTA, a window-specific encoder E_r is paired with a start-space decoder D_{a_r}. Both endpoint codes decode to the same window input:

$$\mathcal{L}_{\mathrm{DA}}^{(r)} = \mathbb{E}\big[\|D_{a_r}(E_r(h_{a_r})) - h_{a_r}\|_2^2 + \|D_{a_r}(E_r(h_{b_r})) - h_{a_r}\|_2^2\big]. \tag{12}$$

The second term is deliberate. It closes the cross-layer transformation by returning the endpoint code to the window-input space. The decoder is used for offline analysis and training of the chart; online LTA uses the frozen encoder only.

For fixed causal context c and a frozen Transformer window Φ_c, define $F = E_r \circ \Phi_c$, $D = D_{a_r}$, and

$$g = F \circ D, \qquad \mathcal{R} = D \circ F. \tag{13}$$

Here g acts in the latent chart and $\mathcal{R}$ acts in the window-input space. No global inverse for either E_r or D is assumed.

Proposition B.1 (Fixed-point correspondence)**.** *If $g(z^\star) = z^\star$, then $h^\star = D(z^\star)$ satisfies $\mathcal{R}(h^\star) = h^\star$. Conversely, if $\mathcal{R}(h^\star) = h^\star$, then $z^\star = F(h^\star)$ satisfies $g(z^\star) = z^\star$.*

Proof. If $g(z^\star) = z^\star$, then $\mathcal{R}(D(z^\star)) = D(F(D(z^\star))) = D(g(z^\star)) = D(z^\star)$. The reverse direction follows from $g(F(h^\star)) = F(D(F(h^\star))) = F(\mathcal{R}(h^\star)) = F(h^\star)$. □

Proposition B.2 (Trajectory correspondence)**.** *For every integer $k \geq 0$,*

$$D \circ g^k = \mathcal{R}^k \circ D, \qquad F \circ \mathcal{R}^k = g^k \circ F. \tag{14}$$

Proof. The one-step identities $D \circ g = \mathcal{R} \circ D$ and $F \circ \mathcal{R} = g \circ F$ follow directly from Eq. (13). Induction gives Eq. (14). □

Proposition B.3 (Local spectral correspondence)**.** *Let $z^\star$ and $h^\star = D(z^\star)$ be corresponding fixed points. If F and D are differentiable there, the Jacobians $J_g(z^\star)$ and $J_{\mathcal{R}}(h^\star)$ have the same nonzero eigenvalues, including algebraic multiplicity. Consequently they have the same spectral radius, apart from possible additional zero eigenvalues caused by unequal dimensions.*

Proof. Writing $A = J_F(h^\star)$ and $B = J_D(z^\star)$ gives $J_g(z^\star) = AB$ and $J_{\mathcal{R}}(h^\star) = BA$. The standard AB–BA spectral identity establishes equality of all nonzero eigenvalues. □

B.1 Exact and approximate transition–field identities

Let $h_b = \Phi_c(h_a)$, $z_a = E_r(h_a)$, and $z_b = E_r(h_b) = F(h_a)$. Under exact dual-anchor reconstruction, $D(z_a) = D(z_b) = h_a$, so

$$g(z_a) = z_b, \qquad g(z_b) = z_b, \qquad V(z_a) = g(z_a) - z_a = z_b - z_a. \tag{15}$$

Thus the realized latent transition equals the residual-field vector at its start, and the observed endpoint is a fixed point of the induced map.

For the approximate case, define

$$\eta_a = \|D(z_a) - h_a\|_2, \qquad \eta_b = \|D(z_b) - h_a\|_2. \tag{16}$$

If F is locally L-Lipschitz around h_a, then

$$\|V(z_a) - (z_b - z_a)\|_2 \leq L\eta_a, \tag{17}$$

$$\|g(z_b) - z_b\|_2 \leq L\eta_b. \tag{18}$$

If D is also L_D-Lipschitz, then

$$\|\mathcal{R}(D(z)) - D(z)\|_2 \leq L_D\|g(z) - z\|_2. \tag{19}$$

The bounds follow by inserting $z_b = F(h_a)$ and applying the respective Lipschitz inequalities. They explain why the two anchors in Eq. (12) control different approximation errors.

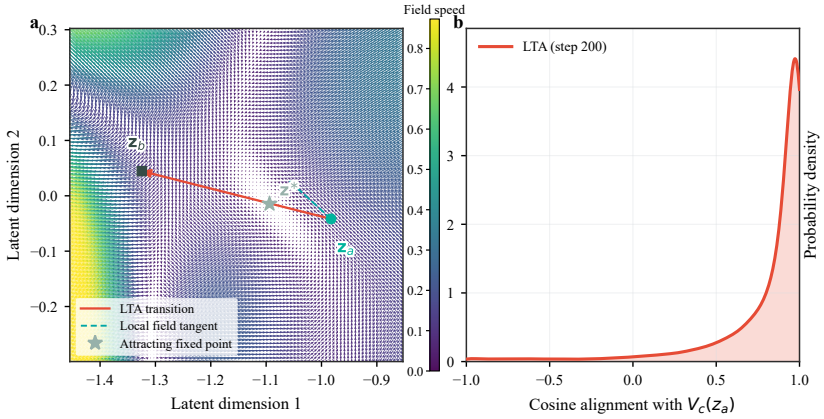

Figure 9: **Controlled local transition–field diagnostic.** The left panel compares the realized latent transition with the fitted residual-field vector for a representative held-out token (cosine 0.882). The right panel shows the local fixed point and registered perturbation returns. The shared frozen observation chart is used for visualization; this diagnostic is distinct from the dual-anchor chart used by online LTA.

For the observational chart, let e_{fit} bound the residual-head fit error on the decoded start. The same argument gives

$$\|\widehat{V}(z_a) - (\widetilde{z}_b - \widetilde{z}_a)\|_2 \le e_{\mathrm{fit}} + L\eta_a. \tag{20}$$

Equation (20) is an observational diagnostic; Eqs. (17)–(18) describe the formal dual-anchor LTA construction.

B.2 Local attraction and persistence

A small fixed-point residual is not sufficient to establish attraction. We therefore separate a pointwise linear test, $\rho(J_g(z^\star)) < 1$, from the stronger uniform contraction condition on a neighborhood. If a closed, convex, forward-invariant set K satisfies $\sup_{z\in K} \|J_g(z)\|_2 \le \kappa < 1$, then g is a contraction on K and has one attracting fixed point there.

The under-relaxed map

$$g_\xi = (1-\xi)\,\mathrm{Id} + \xi g, \qquad 0 < \xi \le 1, \tag{21}$$

has the same fixed points as g. If g is κ-contractive, then g_ξ is at most $1-\xi(1-\kappa)$-contractive. Under-relaxation can therefore stabilize iteration without changing the solutions being examined.

Theorem B.4 (Persistence of a locally attracting structure)**.** *Let $K = \overline{B}(z^\star, r)$, let $g_{\mathrm{ref}}(z^\star) = z^\star$, and assume $\sup_K \|Jg_{\mathrm{ref}}\|_2 \le \kappa < 1$. Suppose*

$$\sup_K \|g_\theta - g_{\mathrm{ref}}\|_2 \le \delta < (1-\kappa)r, \qquad \sup_K \|Jg_\theta - Jg_{\mathrm{ref}}\|_2 \le \eta < 1-\kappa. \tag{22}$$

Then g_θ maps K into itself, is a contraction on K, and has a unique attracting fixed point $z_\theta^\star \in K$ satisfying

$$\|z_\theta^\star - z^\star\|_2 \le \frac{\delta}{1-\kappa}. \tag{23}$$

Proof. For $z \in K$, $\|g_\theta(z) - z^\star\|_2 \le \delta + \kappa r < r$, so $g_\theta(K) \subset K$. Moreover $\sup_K \|Jg_\theta\|_2 \le \kappa + \eta < 1$. Banach's fixed-point theorem gives the unique attracting fixed point. Finally, $\|z_\theta^\star - z^\star\|_2 \le \delta + \kappa\|z_\theta^\star - z^\star\|_2$, which yields the stated bound. □

The local Tiny-NanoChat audit uses all three empirical gates: residual, Jacobian spectral radius, and perturbation return. Across 32,768 held-out tokens, transition–field cosine has median 0.940 and IQR 0.782–0.988. The displayed fixed point has residual 4.5×10^{-7} and spectral radius 0.933; every registered perturbation returns. These measurements support a local mechanism and are not a claim of global autonomous Transformer dynamics.

C Online Objective, Geometry, and Gradients

For every valid token, the online transitions are

$$\Delta z^{\mathrm{ref}}_{r,u} = \mathrm{sg}\big[E_r(h^{\mathrm{ref}}_{b_r,u}) - E_r(h^{\mathrm{ref}}_{a_r,u})\big], \tag{24}$$

$$\Delta z^{\theta}_{r,u} = E_r(h^{\theta}_{b_r,u}) - E_r(h^{\theta}_{a_r,u}), \tag{25}$$

where the reference branch and all surrogate parameters are detached. The current transition retains gradients through the frozen encoder to the trainable hidden states. For d_z latent dimensions,

$$v(a) = \frac{\sqrt{\|a\|_2^2 + \varepsilon}}{\sqrt{d_z}}, \qquad w(b) = \exp[-\beta v(b)^2], \tag{26}$$

and the implementation uses

$$\cos_\varepsilon(a,b) = \frac{\langle a, b\rangle}{\max(\|a\|_2, \varepsilon)\max(\|b\|_2, \varepsilon)}. \tag{27}$$

With valid-token mask μ_u and $N = \sum_u \mu_u$, the two terms are

$$\mathcal{L}^{(r)}_{\mathrm{dir}} = \frac{\lambda_{\mathrm{dir}}}{N}\sum_u \mu_u w^{\mathrm{ref}}_{r,u}\left[1 - \cos_\varepsilon(\Delta z^{\theta}_{r,u}, \Delta z^{\mathrm{ref}}_{r,u})\right], \tag{28}$$

$$\mathcal{L}^{(r)}_{\mathrm{mag}} = \frac{\lambda_{\mathrm{mag},r}}{N}\sum_u \mu_u w^{\mathrm{ref}}_{r,u}\left[v(\Delta z^{\theta}_{r,u}) - \gamma v(\Delta z^{\mathrm{ref}}_{r,u})\right]_+^2. \tag{29}$$

The full loss is

$$\mathcal{L} = \mathcal{L}_{\mathrm{LM}} + \alpha \sum_{r\in\mathcal{W}} \omega_r(\mathcal{L}^{(r)}_{\mathrm{dir}} + \mathcal{L}^{(r)}_{\mathrm{mag}}). \tag{30}$$

Feasible-set geometry. If both current endpoints receive the same latent translation, their transition is unchanged, so LTA does not penalize the shift. If $\Delta z^{\theta} = \kappa \Delta z^{\mathrm{ref}}$ for $\kappa > 0$, the angular penalty is zero away from the stabilizer. Contractions and expansions within the margin are also unpenalized by the one-sided magnitude term. By contrast, StateMatch penalizes both translated endpoints, while Transition-MSE has a unique zero at exact vector equality. For angle ϕ between nonzero transitions, $1-\cos\phi = 2\sin^2(\phi/2)$, making the direction penalty locally quadratic in small angular error.

Piecewise direction gradient. Let $f(a,b) = 1 - \cos_\varepsilon(a,b)$ and $B = \max(\|b\|_2, \varepsilon)$. Away from the switching boundary,

$$\nabla_a f(a,b) = \begin{cases} -\dfrac{b}{\|a\|_2 B} + \dfrac{\langle a,b\rangle a}{\|a\|_2^3 B}, & \|a\|_2 > \varepsilon, \\ -\dfrac{b}{\varepsilon B}, & \|a\|_2 < \varepsilon. \end{cases} \tag{31}$$

At $\|a\|_2 = \varepsilon$, the implementation follows the automatic differentiation branch selected by the maximum operation.

Magnitude gradient and activation boundary. Let $\xi = v(a) - \gamma v(b)$. For an active hinge, the gradient of one weighted term is

$$2\lambda w \xi \frac{a}{\sqrt{d_z}\sqrt{\|a\|_2^2 + \varepsilon}}, \tag{32}$$

and it is zero when $\xi \le 0$. The exact activation boundary is

$$\|a\|_2^2 = \gamma^2 \|b\|_2^2 + (\gamma^2 - 1)\varepsilon. \tag{33}$$

The reference-dependent weights are detached, so they cannot be reduced by changing the trainable branch.

D Implementation and Numerical Contract

D.1 Precision, masking, and graph isolation

Forward and backward Transformer passes use BF16. Encoded transitions, weights, structural losses, and their reductions use FP32. Padding and invalid causal positions are removed before the masked mean. Only the frozen reference branch may be cached; current transitions are recomputed so that their graph remains connected to θ. The two windows contribute equally. In the registered parameterization, $\alpha = 1.105150686754002$, $\rho = 6579.538862439464$, $\beta = 0.3$, $\gamma = 1.0175655364990235$, and $\varepsilon = 10^{-8}$. The corresponding effective direction and magnitude coefficients are 22.10301373508004 and 72713.81892349619 after the base loss scales are included. StateMatch and Transition-MSE use fixed structural scales 0.6366986369259453 and 8.755994913208567, determined by real-batch gradient matching before formal outcomes were inspected.

Table 1: Frozen formal Qwen training configurations. The 3B experiment inherits the loss and optimizer values without a scale-specific search.

Item	Qwen2.5-1.5B	Qwen2.5-3B
Windows	$2\to6$, $14\to18$	$3\to8$, $18\to23$
Latent width	128 per window	128 per window
Optimizer	AdamW, no scheduler	AdamW, no scheduler
Learning rate / weight decay	4×10^{-5} / 0.01	same
Betas / Adam epsilon	$(0.9, 0.999)$ / 10^{-8}	same
Sequence length	512	512
Physical / accumulated batch	4 / 8	4 / 8
Effective batch	32 sequences	32 sequences
Gradient-norm ceiling	30	30
Training horizon	300 updates per domain	300 updates
Evaluation grid	$0, 20, \ldots, 300$	$0, 50, \ldots, 300$

D.2 One replay-free LTA update and numerical gates

For each current-domain batch, the registered runner performs the following sequence:

1. evaluate the rolling reference without gradients and retain only the two endpoint states for each selected window;
2. evaluate the trainable model on the same tokens and compute the language-model loss;
3. encode matched FP32 endpoint pairs, form current and reference transitions, apply the frozen token mask and reference-dependent weights, and accumulate the direction and magnitude terms;
4. backpropagate Eq. (30), clip the global gradient norm, and take one AdamW step.

Only the frozen-reference branch may be cached. The current transitions are always recomputed, so their graph remains connected to the trainable model. Before formal training, independent forward graphs test direction-only, magnitude-only, and complete structural losses below, on, and above the magnitude boundary. The same gate checks mask exclusion, a near-zero reference transition, and a real base-versus-base batch. Gradients are cleared between cases; no optimizer step is taken during this audit.

At every domain boundary in sequential learning, the trained model becomes the next stage's initialization and a frozen copy becomes its rolling reference; optimizer state is reset. Surrogates are frozen throughout continual learning and discarded after training, adding no inference-time parameters or latency.

E LAYER-WINDOW SELECTION AND DEVELOPMENT PROTOCOL

E.1 SPECTRAL ENTROPY DEFINITIONS

For a matrix $X \in \mathbb{R}^{n \times d}$ with singular values $\{\sigma_j\}$, define

$$p_j = \frac{\sigma_j^2}{\sum_k \sigma_k^2}, \qquad H(X) = -\sum_j p_j \log p_j. \tag{34}$$

The effective rank is $\exp H(X)$. Lower entropy means that representational energy is concentrated in fewer dominant directions; it is not an estimator of task accuracy, semantic content, or mutual information.

For the *dataset entropy*, valid token states are mean-pooled within each prompt and the resulting $N \times d_h$ matrix is passed to Eq. (34) without centering. The *mean per-prompt token entropy* applies the same uncentered definition to each prompt's valid-token matrix $T_i \times d_h$ and then averages across prompts. The *centered dataset entropy* subtracts the columnwise mean from the pooled $N \times d_h$ matrix before applying the same equation; it is a sensitivity analysis rather than a replacement metric. Curves in the main paper are divided by each curve's value at state 3. The three definitions agree on the frozen model's compression–re-expansion organization and motivate the 1.5B windows $2 \to 6$ and $14 \to 18$. The 3B windows $3 \to 8$ and $18 \to 23$ transfer these locations by relative depth and pass frozen finite-value and nonzero-movement checks.

E.2 METHOD-BLIND WINDOW SELECTION

The depth diagnostic uses only the frozen base model and input text. It is computed before any continual-learning method is run and is therefore not an outcome-based window search. Finance, Chemistry L3, and MedMCQA exhibit the same broad compression–re-expansion organization under the registered definition. The diagnostic identifies admissible regions rather than a unique causal optimum.

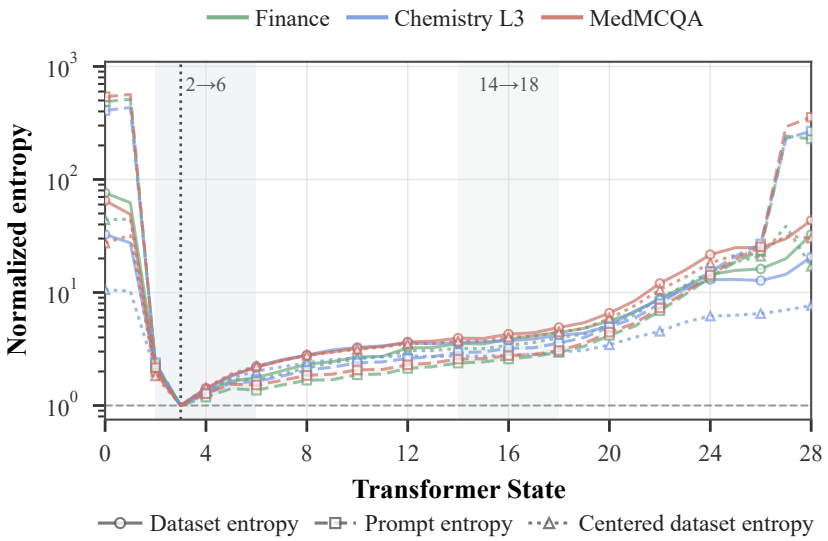

Figure 10: **Frozen depth diagnostic used to delimit candidate windows.** Curves are computed from frozen hidden states before the formal method comparison. The shaded regions mark the early and late admissible windows. Entropy is a compression landmark and is not interpreted as task accuracy or semantic content.

Within those regions, a Finance-only development audit compared early, middle, long-chord, and dual-window variants at a fixed budget. The dual window was the only candidate with a favorable

adaptation–retention relation and was frozen before the formal Chemistry, QASC, and ReClor outcomes were inspected. For Qwen2.5-3B, the corresponding locations were transferred by relative depth rather than re-selected using QASC results.

E.3 Separation of hyperparameter development from formal evaluation

The direction–magnitude form was calibrated on PubMed development data and cs.CL cross-domain validation. Two candidates then underwent a three-seed, 250-step paired confirmation. The selected setting retained 98.85% of SFT adaptation on average while reducing transition drift by 60.44%. Only after the configuration was frozen was Mathematics opened; there it retained 99.07% of SFT adaptation and reduced transition drift by 60.65%. Mathematics did not change the configuration. Learning-rate and clipping audits were kept separate from the structural-coefficient search. Raising the gradient-norm ceiling from 1 to 30 removed a nearly universal clipping regime; the formal experiments therefore use the ceiling reported in Table 1 without revisiting the LTA coefficients.

F Surrogate Training and Held-Out Gates

F.1 Data separation and frozen fitting

All surrogates are fitted before continual learning using a frozen base model and are then immutable. Neither formal target outcomes nor hidden test data enter their fitting. For Qwen2.5-1.5B, the frozen base model extracts 64 C4 development sequences of length 128. The first 48 sequences form the optimization stream and the final 16 are held out. Both windows use the same sequences. For Qwen2.5-3B, both windows use the first 64 examples of a frozen ReClor training stream at length 128, with the same 48/16 split; the resulting surrogates are reused unchanged in the QASC experiment.

F.2 Architecture and optimization

The Qwen encoder is $d_h \to 256 \to 128$ and the decoder is $128 \to 256 \to d_h$, with GELU nonlinearities. Training uses Eq. (12) only: no labels, language-model loss, contrastive term, or downstream metric is present. Optimization uses FP32 AdamW, learning rate 3×10^{-4}, weight decay 10^{-4}, token-pair batch size 4096, and gradient-norm ceiling 1. Every 1.5B and 3B window uses its 3000-step surrogate checkpoint, matching the protocol stated in the main paper.

F.3 Normalization and held-out acceptance gates

After fitting, the latent chart is centered by the optimization-stream mean and divided by one global root-mean-square scale. The decoder applies the inverse normalization before reconstructing the start space. Candidate checkpoints must have finite held-out loss, positive R^2 at both anchors, at least 1% joint-error improvement, non-collapsed latent variance, nonzero median movement, and at least 80% valid-movement coverage. The selected 1.5B early chart has held-out start/end R^2 of $0.9981/0.9979$. At 3B, the early and late charts obtain $0.9978/0.9975$ and $0.9896/0.9892$, respectively, with movement coverage 1.0. These are chart-quality gates, not downstream evidence for LTA.

G Datasets, Selection, and Evaluation Protocol

G.1 TinySciFiction construction

TinySciFiction is a paired, structure-matched adaptation stream derived from TinyStories (Eldan & Li, 2023). Batched DeepSeek V4 rewrites retain the event and discourse structure of each original story while replacing fairy-tale entities, settings, and vocabulary with simple science-fiction counterparts. Records missing either side are removed. The 7,874 remaining pairs are shuffled once with seed 42; the first 374 pairs form validation and the remaining 7,500 form training. Training uses only the science-fiction side. New-domain NLL is evaluated on science-fiction validation text,

while paired forgetting is evaluated on the corresponding original text. The resulting Parquet files contain one text column and row groups of 64.

Table 2: TinySciFiction assets after the frozen construction procedure.

Asset	Records	Approximate size
Raw paired JSONL	7,874	14.97 MB
Science-fiction training split	7,500	4.06 MB
Science-fiction validation split	374	0.21 MB
Paired original validation split	374	0.22 MB

G.2 Formal Qwen domains and references

The formal 1.5B tasks are public development data from SciKnowEval Chemistry L3 (Feng et al., 2024), QASC (Khot et al., 2020), and ReClor (Yu et al., 2020); test sets are not accessed. These domains were selected by a method-blind SFT-only screen requiring both measurable acquisition and measurable reference drift. They share a multiple-choice evaluation interface while spanning scientific knowledge, multi-step scientific reasoning, and logical reasoning. No KL or LTA outcome entered selection.

Table 3: Frozen target-domain streams and development evaluation manifests.

Dataset	Primary emphasis	Training stream	Development evaluation	Options
Chemistry L3	scientific knowledge	1,682 unique examples; deterministic cycling	300-example curve subset; 363-example full check	4
QASC	multi-step scientific reasoning	16,000 materialized positions	600-example curve subset; 926-example full check	8
ReClor	logical reasoning	16,000 materialized positions	500 development examples	4

Within each domain and seed, all methods receive the same initialization, ordered batches, token masks, and evaluation examples. Target learning is measured by development accuracy. Reference forgetting is the change in token NLL relative to the frozen stage input; positive $\Delta\mathrm{NLL}$ means degradation. For the 3B experiment, C4 development, WikiText validation, and a fixed general-reference mixture are differenced separately against step 0 and then averaged within each trajectory. None of these reference examples is replayed during training.

G.3 Method-blind domain suitability screen

Before committing the formal matched comparisons, an SFT-only screen was used to identify streams with both a measurable acquisition signal and measurable reference drift. No Output-KL, StateMatch, Transition-MSE, or LTA result entered this screen. For task data, acquisition was measured by accuracy or macro-F1; for corpus data, it was measured by target-NLL reduction. Most screening cells used one ordered-data seed and are therefore treated as exploratory selection evidence rather than as a method comparison.

Seventeen domains passed the registered screen. PubMed RCT20k provided the strongest confirmed task-learning signal: macro-F1 increased from 14.34% to 79.50% at step 100 and 82.42% at step 300, while mean reference drift rose from 0.0844 to 0.1575. For corpus adaptation, M2D2 eess.AS reduced target NLL by 0.2248 at step 100 and 0.2533 at step 300, with corresponding reference drift of 0.1309 and 0.1959. These measurements establish that the candidate streams expose both learning and forgetting; they do not rank preservation methods.

G.4 Directional multi-domain screening and final curriculum control

Because domain compatibility is directional, an exploratory SFT-only audit also evaluated 22 ordered pairs. A complete M2D2 triangle consisting of q-fin.RM, q-bio.NC, and cs.CR passed all

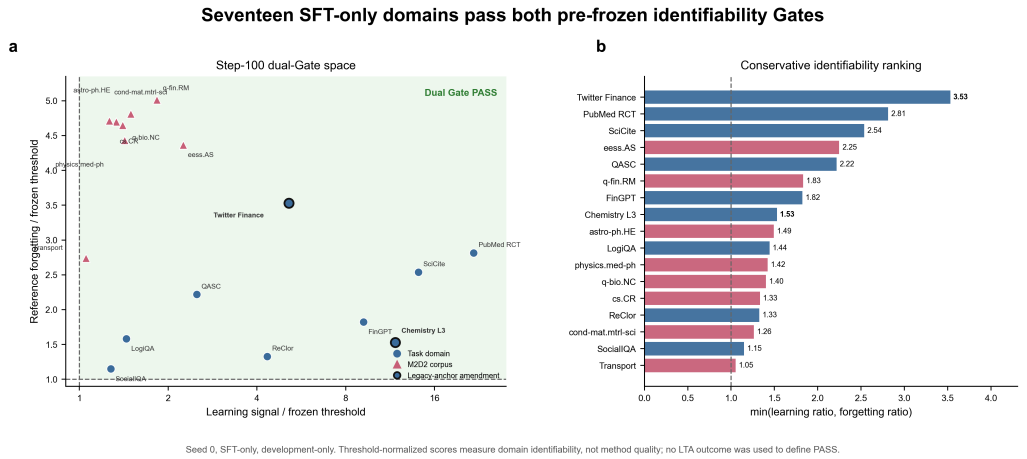

Figure 11: **Method-blind SFT domain triage.** The screen asks whether a candidate stream provides enough acquisition and forgetting signal for a later continual-learning comparison. These SFT-only measurements are not pooled with the five-seed formal results and provide no evidence for LTA.

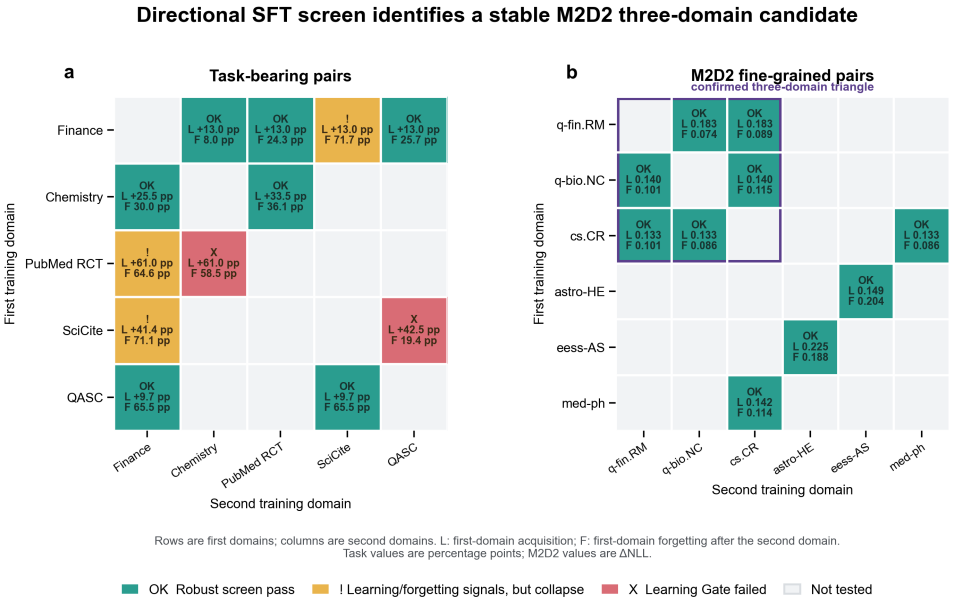

Figure 12: **Directed SFT-only compatibility screen.** Rows denote the source domain and columns the subsequent domain. The audit records acquisition and measurable forgetting in both directions and is used only to diagnose order sensitivity.

six directed edges, followed by a two-seed, 150-step-per-stage confirmation. Its mean acquisition-NLL decrease was 0.1459 and its old-domain backward-transfer ΔNLL was $+0.1114$. This audit established that order alone can materially alter apparent continual-learning difficulty.

The final accuracy-based sequential study instead uses the three task-bearing domains already frozen for the single-domain comparison and evaluates three pre-specified orders. This choice keeps the single-domain and sequential evidence on the same tasks. After an order is fixed, every method and seed receives the same initialization, ordered examples, stage boundaries, and evaluation manifests; no curriculum is retained or discarded on the basis of an LTA outcome.

H Baselines, Ablations, and Comparison Boundary

SFT. The replay-free lower control optimizes current-domain next-token cross-entropy only.

Output KL. Following Learning without Forgetting (Li & Hoiem, 2016), the frozen reference and current model process the same current-domain tokens. We use $D_{\mathrm{KL}}(p_{\mathrm{ref}} \| p_\theta)$ over the full vocabulary at temperature 1 and weight 1 on all valid causal positions; completion masks apply only to the language-model loss.

StateMatch. Both encoded endpoints are matched to the rolling reference:

$$\mathcal{L}_{\mathrm{state}} = \mathcal{L}_{\mathrm{LM}} + \lambda_s \sum_{r,u} \mu_u \big(\| z^{\theta}_{a_r,u} - z^{\mathrm{ref}}_{a_r,u} \|_2^2 + \| z^{\theta}_{b_r,u} - z^{\mathrm{ref}}_{b_r,u} \|_2^2 \big). \tag{35}$$

Transition-MSE. The complete latent transition is matched symmetrically:

$$\mathcal{L}_{\Delta\mathrm{MSE}} = \mathcal{L}_{\mathrm{LM}} + \lambda_\Delta \sum_{r,u} \mu_u \| \Delta z^{\theta}_{r,u} - \Delta z^{\mathrm{ref}}_{r,u} \|_2^2. \tag{36}$$

Direction-only and magnitude-only. These ablations retain the full dual-window normalization and all training controls, but include only the corresponding LTA term. StateMatch, Transition-MSE, and both ablations share the same reference model, windows, encoders, masks, batches, learning rate, and training budget with LTA.

FDD (Gong et al., 2025) projects layer-wise changes through language-model heads and distills feature trajectories; MTA (Chi et al., 2026) uses a teacher–student design across word-, phrase-, and structure-level representations. Function-vector guidance (Jiang et al., 2025) and RECALL (Wang et al., 2025) use different intervention or model-combination resources. These works motivate the comparison boundary but are not treated as matched replay-free baselines because their available resources and update objects differ from the single rolling-reference protocol used here.

I Complete Tables for Results Reported in the Main Paper

This section tabulates the endpoint quantities discussed in the main paper. It does not add post-hoc outcome screens or unrelated benchmark diagnostics. Every Qwen entry is the mean over five paired seeds. Values following $\pm$ are 95% Student-t confidence-interval half-widths.

I.1 Controlled TinyStories–TinySciFiction intervention

Table 4: Tiny-NanoChat endpoints at adaptation step 300. ForgetNLL is lower when the original domain is better preserved; NewGain and transition cosine are higher when adaptation and structural alignment are stronger.

Method	ForgetNLL ↓	NewGain ↑	Transition cosine ↑
SFT	1.2447	1.0036	0.3849
LTA intervention	**0.8745**	**1.2610**	**0.7452**

The intervention reduces ForgetNLL by 29.75%, while increasing rather than reducing the new-domain NLL gain. The aligned transition cosine almost doubles, connecting the retention improvement to the structural quantity targeted by the intervention.

I.2 SINGLE-DOMAIN QWEN2.5-1.5B RESULTS

Table 5: Single-domain endpoints after 300 updates. Accuracy is measured on the corresponding development set and reference ΔNLL against the frozen base model. Bold marks the best value in each column.

(a) Development accuracy (%)

Method	Chemistry	QASC	ReClor
SFT	74.07±2.37	75.23±2.19	70.92±.83
Output KL	**74.52±1.42**	**77.10±1.01**	70.72±1.94
StateMatch	73.38±1.20	67.27±7.87	68.40±2.13
Transition-MSE	73.54±1.91	74.63±3.39	**71.16±2.33**
Direction	74.22±1.68	75.47±1.39	67.92±1.44
Magnitude	74.30±2.84	76.27±2.30	69.68±2.13
LTA	74.14±1.20	76.27±1.88	68.16±1.38

(b) Reference ΔNLL

Method	Chemistry	QASC	ReClor	Mean
SFT	.0714±.0109	.0961±.0162	.0808±.0155	.0828
Output KL	.0239±.0076	.0392±.0074	.0169±.0065	.0267
StateMatch	.0488±.0137	.1021±.0620	.0525±.0140	.0678
Transition-MSE	**.0227±.0092**	.0484±.0190	.0221±.0045	.0310
Direction	.0250±.0061	.0342±.0092	.0071±.0022	.0221
Magnitude	.0626±.0197	.0805±.0197	.0573±.0108	.0668
LTA	.0269±.0158	**.0319±.0078**	**.0070±.0050**	**.0220**

Full LTA has the lowest three-domain mean drift. Relative to Output KL, Transition-MSE, and SFT, its mean drift is lower by 17.6%, 29.2%, and 73.5%, respectively. Direction-only is the closest component ablation; magnitude-only is insufficient. On ReClor, Transition-MSE has higher target accuracy, but its reference drift is more than three times LTA's.

I.3 SEQUENTIAL CONTINUAL LEARNING ACROSS THREE ORDERS

Let $A_{1,1}$ be first-domain accuracy after the first stage and $A_{1,3}$ the same accuracy after all three stages. Accumulated forgetting is $F_1 = A_{1,1} - A_{1,3}$; lower is better.

Table 6: First-domain retention under three sequential orders. All values are percentage points. Bold marks the best final retention and lowest forgetting in each order.

Order (retained domain)	Method	$A_{1,1}$	$A_{1,3}$	$F_1 \downarrow$
QASC→Chemistry→ReClor (QASC)	SFT	75.20±1.83	42.60±5.77	32.60
	Output KL	77.10±1.14	43.83±5.78	33.27
	StateMatch	67.27±7.92	37.27±2.81	30.00
	Transition-MSE	74.50±3.45	42.37±5.97	32.13
	LTA	76.20±1.86	**51.13±8.22**	**25.07**
ReClor→Chemistry→QASC (ReClor)	SFT	70.04±1.84	63.04±1.80	7.00
	Output KL	70.84±2.01	64.20±2.16	6.64
	StateMatch	68.76±3.82	63.36±2.43	5.40
	Transition-MSE	70.08±2.15	64.16±2.30	5.92
	LTA	68.12±1.54	**64.48±1.60**	**3.64**
ReClor→QASC→Chemistry (ReClor)	SFT	70.04±1.84	62.72±2.33	7.32
	Output KL	70.84±2.01	63.52±1.99	7.32
	StateMatch	68.76±3.82	62.32±4.09	6.44
	Transition-MSE	70.76±2.20	63.36±3.06	7.40
	LTA	68.12±1.54	**66.16±1.13**	**1.96**

LTA has the highest final first-domain accuracy in all three orders. Relative to Output KL, accumulated forgetting is lower by 8.20, 3.00, and 5.36 points, respectively. The conclusion therefore does not depend on one favorable placement of Chemistry or QASC after the retained task.

I.4 QWEN2.5-3B SCALE TRANSFER

Table 7: Qwen2.5-3B QASC endpoints. Aggregate drift is the within-run mean over C4, WikiText, and the fixed general-reference set. Per-reference entries are reported as mean [95% confidence interval].

(a) Target accuracy and aggregate drift

Method	Accuracy (%)	Aggregate ΔNLL
SFT	78.90±1.52	.0666±.0133
Output KL	**78.77±1.65**	.0370±.0071
StateMatch	77.27±2.46	.0447±.0114
Transition-MSE	78.43±1.38	.0293±.0071
LTA	77.17±3.99	**.0108±.0029**

(b) Per-reference ΔNLL

Method	C4	WikiText	General
SFT	.0675 [.0577, .0773]	.0784 [.0537, .1031]	.0539 [.0389, .0688]
Output KL	.0533 [.0454, .0612]	.0308 [.0206, .0410]	.0268 [.0198, .0339]
StateMatch	.0548 [.0388, .0707]	.0487 [.0363, .0611]	.0305 [.0201, .0410]
Transition-MSE	.0340 [.0268, .0412]	.0355 [.0283, .0428]	.0183 [.0087, .0279]
LTA	**.0221** [**.0190**, **.0252**]	**.0128** [**.0100**, **.0156**]	**−.0025** [**−.0069**, **.0018**]

At step 300, LTA's aggregate drift is 63.1% below Transition-MSE, the strongest non-LTA retention baseline. LTA reaches 77.17% QASC accuracy, a 3.33-point gain over the frozen base model's 73.83%. Output KL attains the highest target accuracy, while its aggregate drift is 3.42 times LTA's. The slightly negative LTA mean on the general-reference subset does not imply a negative aggregate: the registered aggregate over all three references is 0.0108.

J CHECKPOINT-ONLY GENERAL-BENCHMARK DIAGNOSTIC

The development study additionally evaluated retained Finance-adaptation checkpoints on MMLU, MMLU-Pro, HellaSwag, GSM8K, IFEval, and MBPP-Sanitized. No additional training, coefficient selection, or checkpoint modification was performed. This audit is retained as a diagnostic module, but its raw score table is not reproduced here because it is not part of the main paper's evidence set.

The six tasks combine knowledge, reasoning, instruction following, exact response syntax, and executable code. Under a small evaluation cap, some scores have coarse resolution, and exact-match tasks can confound retained capability with response-schema interference. In the Finance audit, for example, adaptation toward a short class-label schema altered parse coverage on strict-generation tasks. Averaging such outcomes would therefore create a scientifically ambiguous "general ability" number. This observation motivates the paper's primary use of frozen, target-independent reference NLL: it is dense, format-agnostic, and defined at every checkpoint. The benchmark audit is neither used for model selection nor presented as evidence of method superiority.

K OPTIMIZATION, RUNTIME, AND STORAGE

K.1 ONLINE COST

All replay-free preservation methods require the trainable forward and backward pass. Output KL additionally retains full-vocabulary teacher distributions, whereas LTA retains only selected hidden-state endpoints and evaluates the frozen encoders. In the registered RTX 5090 cadence used for the controlled runtime audit, normalized single-domain training time was 1.00× for SFT, 1.28× for single-window LTA, and 1.73× for Output KL. This comparison excludes the one-time surrogate

fitting cost, because a frozen surrogate is reused across downstream domains. These ratios are hardware- and implementation-dependent and are not extrapolated to every accelerator.

The final dual-window 1.5B audit required approximately 15–17 minutes per 300-step SFT cell and 21–23 minutes for Output KL or LTA. In the separate 3B probe, uninterrupted cells required approximately 31.2 minutes for Output KL and 34.3 minutes for LTA. Peak allocated memory in that probe was 29.62 GiB for SFT, 35.92 GiB for Output KL, and 37.37 GiB for LTA. These measurements describe the registered implementations rather than an asymptotic complexity claim.

K.2 Checkpoint retention and recovery boundary

Model-only states are retained at the analysis points needed by the paper. Full optimizer state is a temporary strict-resume asset rather than a reported model artifact. A valid resume requires the same method configuration, model and optimizer states, random-number-generator state, data cursor, and next ordered batch. A checkpoint from one method, seed, learning rate, or domain never initializes another independent cell. Sequential learning is the only intended exception: each stage starts from the preceding stage's trained model and resets the optimizer, as specified in Section D.

L Statistical Aggregation

For a five-seed quantity y_i, figures and tables report

$$\bar{y} \pm t_{0.975,4}\frac{s}{\sqrt{5}}, \qquad t_{0.975,4} = 2.776, \tag{37}$$

where s is the sample standard deviation. Curves are aggregated only after matching method, seed, step, domain, and evaluation state. In the 3B aggregate, reference changes are first computed relative to step 0 within each seed, then averaged across the three reference sets, and only then aggregated across seeds. The single-domain macro drift is likewise computed from the three domain endpoints for each method. Tiny-NanoChat bands in the main paper are sample standard deviations over three seeds, as stated in its caption.

Every formal run records the method configuration, training history, evaluation rows, and completion status. Independent cells always start from the original base checkpoint; only the intended sequential stages inherit the previous stage's model. No failed seed is silently removed, curves are not smoothed, and data order is not replaced on the basis of outcomes.

M Reproducibility Map

The appendix separates the information needed to reconstruct each stage of the evidence chain. Table 8 summarizes where the corresponding contract is specified.

Table 8: Reproducibility map for the reported experiments.

Component	Recorded information
Latent construction	endpoint convention, chart architecture, normalization, and held-out gates
Online objective	masks, detached reference quantities, coefficients, precision, and gradients
Data construction	source datasets, split discipline, deterministic stream construction, and evaluation manifests
Formal training	initialization, ordered batches, optimizer, stage boundary, seed, and completion status
Evaluation	target metric, reference sets, checkpoint grid, and within-seed differencing rule
Aggregation	paired seed set, Student-t interval, and order of averaging operations

Together, Sections D–G specify the training and data contracts, while Sections I and L specify the reported quantities and aggregation. The formal run records preserve configurations, training

histories, evaluation rows, and explicit completion states. Test splits are not introduced, failed seeds are not silently removed, and outcome-dependent data-order substitution is prohibited.

N Limitations and Interpretation Boundary

The five paired seeds quantify variability under the frozen data orders, not under every prompt, model family, or training horizon. Three curricula make the order evidence substantially stronger, but they do not cover all possible domain sequences. The entropy profile is a model-specific compression landmark that narrows window selection; it does not prove that entropy alone identifies a universally optimal transition. Dual-anchor reconstruction is approximate, which is why field alignment, fixed-point residual, Jacobian contraction, and perturbation return are measured separately. The 3B QASC experiment transfers the frozen protocol without scale-specific loss tuning; its superior retention does not imply the highest target accuracy. Finally, LTA adds no inference-time component after training, but online training still requires a frozen-reference forward pass and storage of selected endpoints.

Across the controlled Tiny-NanoChat intervention, three single-domain comparisons, three curricula, and the 3B transfer, the strongest supported conclusion is that preserving selected latent transitions consistently reduces reference drift relative to unconstrained SFT. Full LTA has the lowest single-domain mean drift, the highest final first-domain accuracy in every registered order, and the lowest 3B aggregate reference drift while still learning QASC.