TAMING ATTENTION TO COMPOSE SOFT PROMPT EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models can be adapted into specialized experts through parameter or soft-prompt tuning. Combining independently trained soft-prompt experts without joint retraining is challenging because direct joint readout can misbalance both **the expert group's total influence** and **the allocation among experts**. We propose **TAME** (**T**wo-level **A**ttention **M**ixing of **E**xperts), a training-free, input-conditioned composition method. TAME uses two-level attention to separately control the expert group's total readout and the within-group contribution of each expert while keeping all experts independently available. On GLUE, using the same set of soft-prompt experts, TAME achieves 94.4% mean capability retention, improves direct joint reading by 11.4 percentage points, and outperforms the evaluated context-space composition baselines. In a unified comparison with parameter merging, TAME achieves 1.90 percentage points higher GLUE Macro accuracy despite a lower single-expert reference, matches or exceeds capability retention on mathematics, code, and instruction following, and exhibits **complementary capability-retention patterns** with parameter merging. Overall, TAME provides a competitive training-free route for composing soft-prompt experts.

1 INTRODUCTION

Large language models acquire broad capabilities through pretraining, while task-specific fine-tuning turns them into experts for different needs. More generally, adaptation can produce specialized capabilities or behaviors in at least two forms: parameter updates that modify the model itself, and context-space soft prompts that inject task-related information into a shared frozen backbone. As the number of experts grows, a central challenge is how to make independently trained experts work together in one model without joint retraining or an additional fusion stage.

Parameter-space model merging addresses this challenge by writing multiple expert updates into shared parameters before inference. It offers a direct way to deploy a single merged model, but the updates become coupled inside the shared computation and can cause task-selective capability loss. A model's average score can therefore conceal which capabilities are preserved and which collapse; prior work has shown that some task combinations repeatedly suffer substantial degradation across parameter-merging methods (Cao et al., 2026). Context-space composition follows a different path: it keeps soft-prompt experts independent and reads them jointly during inference (Fig. 1).

Soft prompts provide a lightweight unit of specialization: each expert can be stored as a small set of continuous vectors on top of the same frozen backbone, with substantially fewer trainable parameters than a typical LoRA expert. Existing context-space composition methods use concatenation, fixed structures, hard routing, or learned gates to combine prompt experts (Asai et al., 2022; Li et al., 2025; Hao et al., 2022; Guardieiro et al., 2025; Venkateswaran and Contractor, 2025). However, independently effective soft prompts do not necessarily remain effective when read together. Joint attention must coordinate both the total influence of the expert group relative to the task text and the allocation of contribution among individual experts. When these two quantities are left uncontrolled, direct joint reading can introduce severe composition loss.

We propose TAME (Two-level Attention Mixing of Experts), a training-free method for composing independently trained soft-prompt experts. TAME encodes each soft prompt as a frozen layer-wise key–value (KV) Bank and keeps all Banks available during inference. An input-level attention probe estimates an expert preference profile and selects a reference Bank. TAME then separates

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.

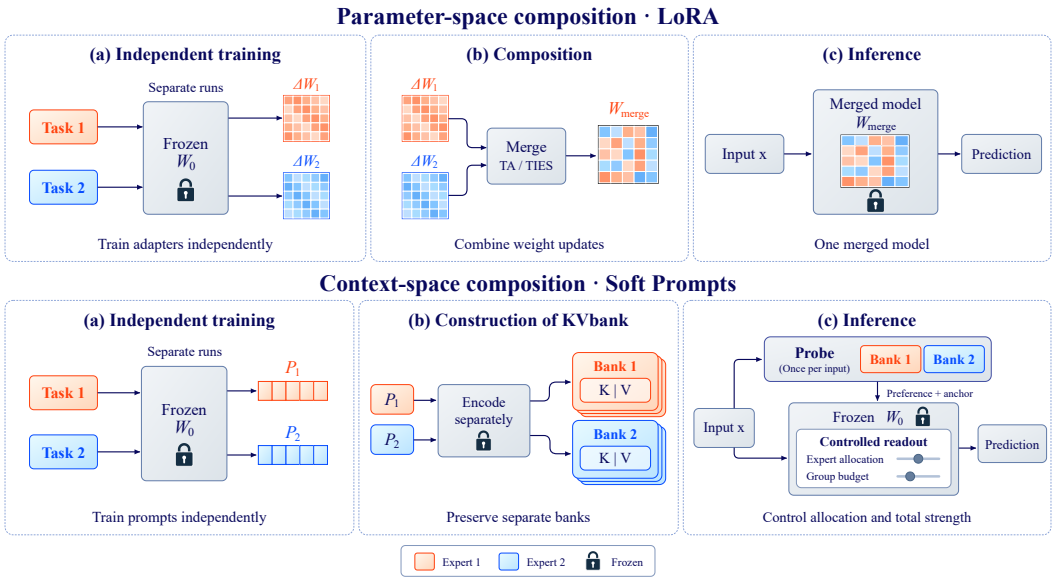

Figure 1: Parameter-space merging versus context-space joint readout. In parameter space, capabilities are merged before inference and remain static during inference. In context space, capabilities are composed during inference through input-adaptive joint readout.

two controls: it first sets the total readout strength of the expert group relative to the task text, and then allocates that budget among experts according to the input preference and the current query. This two-level readout preserves expert independence while allowing their contributions to adapt during inference. TAME requires neither ground-truth task identity at inference nor additional fusion training.

We first compare different ways of using the same set of soft-prompt experts. On the GLUE tasks, TAME reaches a 94.4% mean expert-capability retention, improving over direct joint reading by 11.4 percentage points and achieving the best overall retention among the evaluated context-space composition methods. Ablations show that controlling the total expert-group readout already removes a substantial part of the composition loss, while input-conditioned expert preference allocation provides an additional improvement. The same composition mechanism also transfers to mathematics, code, and instruction-following experts with different soft-prompt lengths.

Finally, we compare TAME with parameter-space model merging to measure the competitiveness of soft-prompt composition and to examine whether the two paths preserve capabilities in the same way. Although the soft-prompt experts have lower average single-expert performance than the LoRA experts on GLUE, their composed result with TAME achieves a higher mean accuracy than LoRA TIES by 1.90 percentage points and substantially reduces the task-level degradation observed in parameter-space merging. On the more demanding mathematics, code, and instruction-following tasks, TAME's mean capability retention matches or exceeds the evaluated parameter-merging methods depending on the preference setting. Finer-grained task and sample analyses show that the two paths preserve different capabilities: parameter merging is stronger on some tasks, while TAME better preserves others that are sensitive to parameter coupling. Soft-prompt composition and parameter-space merging therefore provide parallel and complementary routes for reusing specialized capabilities.

Our key contributions are as follows:

1. We systematically characterize training-free soft-prompt composition and identify two imbalances in direct joint readout: the total readout strength of the expert group and the allocation among experts.
2. We propose TAME, an input-conditioned two-level attention method that controls these factors separately, reducing composition loss while preserving expert independence.

3. We provide the first direct comparison between parameter merging and soft-prompt composition under a unified setting. TAME brings context-space composition to a performance level comparable to parameter merging, and our analysis reveals complementary capability-retention patterns across the two composition paths.

2 Related Work

Model Merging and Task Interference. Parameter merging integrates independently trained updates into a shared model, spanning task-vector arithmetic, conflict resolution via trimming and sign agreement (Ilharco et al., 2023; Yadav et al., 2023), update-space alignment for LoRA experts (Stoica et al., 2025; Gargiulo et al., 2025), and further variants refining update selection, aggregation, and mergeable fine-tuning (Yu et al., 2024; Matena and Raffel, 2021; Marczak et al., 2025; Cheng et al., 2025; Zeng et al., 2025; Zhou et al., 2026). Yet degradation remains task-selective—some tasks persistently lose capability across methods, and common conflict metrics correlate poorly with it (Cao et al., 2026)—and input-dependent expert use also exists on the parameter side (Lu et al., 2024; Lee et al., 2026); our comparison therefore targets the static merges that dominate this line. We accordingly contrast aggregation into shared parameters with joint readout of separately retained contextual experts.

Soft Prompts as Task Experts. Soft prompts adapt frozen models with few trainable parameters: Prompt Tuning optimizes input-layer prompt embeddings, whereas Prefix Tuning trains per-layer prefix representations (Lester et al., 2021; Li and Liang, 2021). Their capability is configuration-dependent but can approach heavier adapters: the vanilla PT baseline in ULPT matches a rank-8 LoRA on Llama-3.2-1B with 20.5K versus 852K trainable parameters (Wu et al., 2026), while theory identifies expressiveness limits of context-based adaptation (Petrov et al., 2024). Prompt experts thus provide a lightweight yet capable unit of capability reuse on a shared frozen backbone.

Contextual Expert Reuse and Attention Readout. Learned prompt mixtures combine prompt experts through jointly trained gates and routing (Asai et al., 2022; Li et al., 2025). Training-free composition spans attention masking and position re-indexing, arithmetic over prompt updates, and KV-carrier composition with readout calibration (Yang et al., 2023; Belanec et al., 2024; Hao et al., 2022; Eyuboglu et al., 2025; Yang et al., 2025). Inference-time steering amplifies attention on user-specified spans (Zhang et al., 2024; Guardieiro et al., 2025; Venkateswaran and Contractor, 2025). These methods either train fusion components or emphasize user-specified spans; TAME instead performs training-free joint readout of independently trained experts under two-level readout control.

3 TAME: Two-level Attention Mixing of Experts

TAME combines independently trained experts through controlled joint readout in context space. Joint readout applies a position reset: expert Banks occupy parallel context positions with no relative order and no attention between Banks, so the downstream text reads them as one parallel expert block. Joint inference introduces attention competition among experts and between the expert group and the input text. TAME separates these interactions into two controls: allocation within the expert group and the group's total readout strength. Like a mixing console, it adjusts individual channels and the master volume separately, allowing multiple experts to contribute while preserving the balance between expert information and task text.

3.1 Problem setup

Each of the T experts owns an input-layer soft prompt trained independently with only that expert present. We encode each prompt through the shared frozen language model into a reusable layer-wise key–value Bank:

$$\mathcal{B}_k = \{(\mathbf{K}_k^{(\ell)}, \mathbf{V}_k^{(\ell)})\}_{\ell=1}^{L}, \tag{1}$$

where L is the number of Transformer layers and each Bank contains M expert positions. Non-expert context comprises the input text, generated history, and a unique shared prefix when present, stored once. The backbone, prompts, and expert Banks remain frozen during composition.

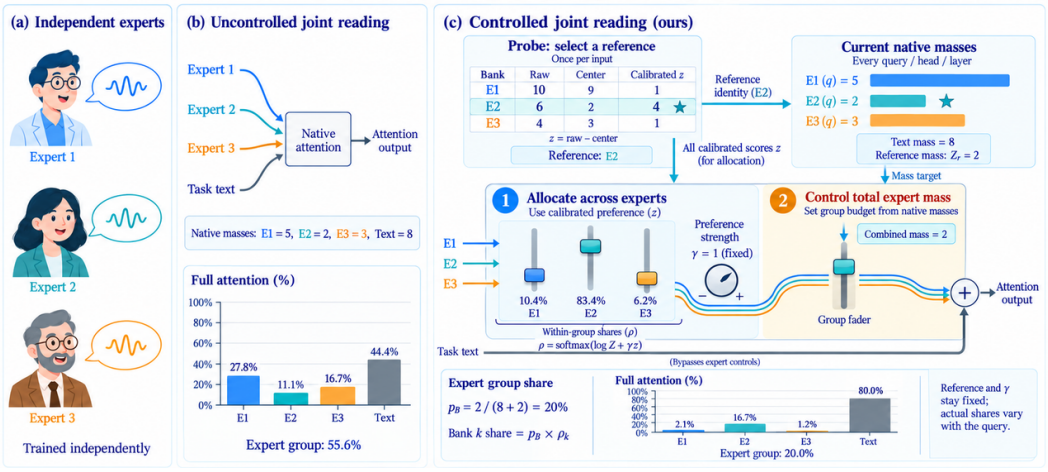

Figure 2: Overview of TAME. Intuitively, TAME is a two-level mixing console: it first reallocates the volume within the expert group, then bounds the group's total volume relative to the task text. (a) Multiple soft-prompt experts are trained independently. (b) Uncontrolled direct joint readout: native attention lets the expert group occupy an uncontrollably high share. (c) TAME's controlled joint readout: each input is probed once, and the calibrated raw scores identify the preferred expert, which serves as the reference Bank. During formal readout, within-group shares are allocated according to the calibrated scores, and the reference Bank's native mass Z_r sets the expert group's total share, while the task-text pathway remains uncontrolled.

3.2 Input-conditioned routed readout

TAME first probes the input under uncontrolled joint attention to estimate expert preferences and select a reference Bank. It then applies two-level control during formal inference, using the reference Bank to set the expert group's readout scale.

Input-level attention score. We probe the last at most four input queries. For an input of length l, let $w = \min(4, l)$ and let $A_{\ell,h,q,kj}(x)$ be the attention probability from query q to position j in Bank k, normalized over all causally visible expert and non-expert positions. The raw Bank score is its attention mass averaged over layers, query heads, and probe positions:

$$\mathrm{raw}_k(x) = \frac{1}{LHw} \sum_{\ell=1}^{L} \sum_{h=1}^{H} \sum_{q=l-w+1}^{l} \sum_{j=1}^{M} A_{\ell,h,q,kj}(x), \tag{2}$$

where H is the number of query heads per layer. The probe runs in evaluation mode.

Mean calibration and reference selection. To remove Bank-specific score baselines, we estimate each Bank's mean on a fixed, unlabeled calibration set of N inputs using the same T-Bank probe:

$$\mu_k = \frac{1}{N} \sum_{i=1}^{N} \mathrm{raw}_k(x_i), \qquad u_k(x) = \mathrm{raw}_k(x) - \mu_k, \qquad r(x) = \arg\max_k u_k(x). \tag{3}$$

The centered score $u_k(x)$ measures preference relative to the Bank's baseline, and the highest-scoring Bank becomes the reference. Calibration statistics μ_k are frozen for inference; ties follow a fixed deterministic rule. Neither probing nor calibration uses task identities or answer labels.

3.2.1 Within-group attention allocation

For the current query, suppressing layer and head indices, let s_{kj} be the original attention logit for position j in Bank k. Define the Bank's native unnormalized mass and its logarithm as

$$Z_k = \sum_j \exp(s_{kj}), \qquad z_k = \log Z_k. \tag{4}$$

Here Z_k captures the local query–Bank match, while $u_k(x)$ supplies the input-level preference. Adding a uniform bias $\gamma u_k(x)$ to each Bank gives

$$\widehat{s}_{kj} = s_{kj} + \gamma u_k(x), \qquad \rho_k = \mathrm{softmax}_k\big(z_k + \gamma u_k(x)\big), \tag{5}$$

where $\gamma \geq 0$ controls preference strength and $\mathrm{softmax}_k$ normalizes over Banks. This acts like adjusting channel levels: each native mass is multiplied by $\exp(\gamma u_k(x))$, while relative token weights within each Bank remain unchanged. The resulting ρ_k specifies each expert's share within the group.

3.2.2 Expert-group mass control

TAME next sets the master level using the reference Bank's native mass Z_r at the current query. A common correction across all Banks preserves their relative shares while setting the group's total unnormalized mass to Z_r:

$$b = z_r - \log \sum_{m=1}^{T} \exp(z_m + \gamma u_m(x)), \qquad \widetilde{s}_{kj} = \widehat{s}_{kj} + b. \tag{6}$$

Non-expert logits remain unchanged. Standard softmax over all visible positions then gives

$$p_{\mathcal{B}} = \frac{Z_r}{Z_{\text{text}} + Z_r}, \qquad p_k = p_{\mathcal{B}} \rho_k, \tag{7}$$

where Z_{text} is the native unnormalized mass of the visible non-expert context. Thus $p_{\mathcal{B}}$ controls the expert group's overall strength, and ρ_k distributes that strength among experts. Setting $\gamma = 0$ gives the mass-only baseline, which retains native Bank proportions while using the probed reference to control total mass.

3.3 Inference procedure

For each input, TAME obtains $\mathbf{u}(x)$ and $r(x)$ from the uncontrolled probe, then reprocesses the input under controlled attention to construct the non-expert context cache. Autoregressive generation reuses the expert Banks and appends new tokens to the non-expert history.

The preferences $\mathbf{u}(x)$, reference identity $r(x)$, and γ remain fixed for one response. Native masses, the common correction, and realized attention shares are recomputed at every layer, query head, and query position. All Banks remain available, allowing their contributions to adapt throughout generation. Algorithm 1 in Appendix A summarizes the full procedure for one input.

4 Experiments

We evaluate TAME as a training-free method for composing independently trained soft-prompt experts on a shared frozen backbone. The experiments are organized around one question: can input-conditioned two-level attention control reduce the capability loss caused by jointly reading multiple context-space experts? We first compare TAME with context-space composition baselines on GLUE, then isolate the contributions of total-mass and within- group control, and finally test the method on more challenging tasks.

4.1 Experimental Setup

Setup. All experiments use Qwen3-1.7B as the frozen base model. Independently trained input-layer soft prompts are encoded as layer-wise key–value (KV) Banks, and composition introduces no additional training. The evaluation covers GLUE, GSM8K, MBPP, and IF Strict Prompt; mathematics and code use 256 virtual tokens, while instruction following uses 128.

We compare TAME with Plain joint, Structured Prompting (Hao et al., 2022), InstABoost (Guardieiro et al., 2025), SpotLight (Venkateswaran and Contractor, 2025), and centered top-1. All methods use the same frozen experts and task-blind inference protocol. The probe and calibration statistics are frozen before inference, with no task identity or answer label supplied online. We report task accuracy, Macro, and mean/worst capability retention relative to the corresponding single context expert.

4.2 Main Results on Soft-Prompt Composition

We first compare ways of jointly reading the same set of soft-prompt experts. This is the primary test of whether TAME improves context-space composition. Table 1 reports absolute accuracy together with the change relative to the corresponding single context expert.

Table 1: Context-space composition on GLUE. Each task cell reports accuracy (%) and the change relative to the corresponding single context-expert reference; color intensity indicates the change.

Method	CoLA	SST-2	QQP	QNLI	RTE	WNLI	Macro	Mean ret.	Worst ret.
Experts (context)	80.82	95.53	88.90	92.35	84.84	50.70	82.19	—	—
Plain joint	70.28 -10.54	89.33 -6.20	63.88 -25.02	71.08 -21.27	49.10 -35.74	56.34 +5.64	66.67	83.0	57.9
Structured Prompting	73.73 -7.09	93.12 -2.41	68.79 -20.11	83.10 -9.25	77.26 -7.58	46.48 -4.22	73.75	89.8	77.4
InstABoost	69.99 -10.83	89.68 -5.85	63.43 -25.47	66.58 -25.77	49.10 -35.74	56.34 +5.64	65.85	82.1	57.9
SpotLight	70.37 -10.45	89.68 -5.85	63.74 -25.16	71.44 -20.91	48.74 -36.10	56.34 +5.64	66.72	83.1	57.4
centered top-1	77.18 -3.64	94.27 -1.26	83.67 -5.23	64.96 -27.39	75.09 -9.75	52.11 +1.41	74.55	91.7	70.3
TAME (mass-only, $\gamma = 0$)	77.47 -3.35	93.35 -2.18	74.81 -14.09	81.20 -11.15	81.59 -3.25	47.89 -2.81	76.05	92.7	**84.2**
TAME ($\gamma = 1$)	77.66 -3.16	93.46 -2.07	74.76 -14.14	81.04 -11.31	83.03 -1.81	52.11 +1.41	**77.01**	**94.4**	84.1

Plain joint drops from the 82.19% single-expert Macro to 66.67%, showing that independently effective Banks are not automatically compatible when read together. Structured Prompting, InstABoost, SpotLight, and centered top-1 recover part of this loss, but each leaves a substantial task-level failure pattern. TAME reaches 77.01% Macro and 94.4% mean retention, with the strongest RTE preservation among the listed context-space methods. The 10.34-point Macro improvement over Plain joint is the direct evidence that the proposed readout control improves soft-prompt composition.

The aggregate metrics also show a useful boundary. Mass-only TAME obtains a slightly higher worst-task retention (84.2%) than the full method (84.1%), while the full two-level method gives the best Macro and mean retention. Thus, the contribution is not a uniform improvement on every metric; it is a better balance between overall capability and retention across tasks.

4.3 Ablations of Two-Level Readout Control

The main table contains a capability ladder from Plain joint to mass-only TAME and then to full TAME. Plain joint amplifies all Banks simultaneously and can over-weight the expert group relative to the task text. Mass-only TAME uses the probe-selected reference to control the group's total readout mass while retaining native Bank proportions. Full TAME adds within-group allocation, allowing the input preference to reweight the experts under the same total-mass budget.

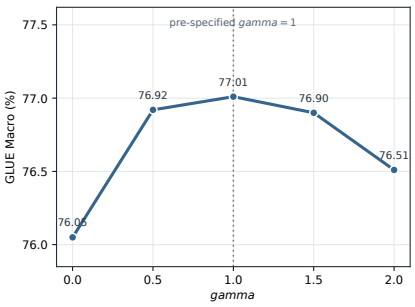

Figure 3: GLUE Macro across γ.

The mass-only row recovers much of the broad degradation of Plain joint, raising Macro from 66.67% to 76.05% and mean retention from 83.0% to 92.7%. The final within-group allocation raises Macro to 77.01% and mean retention to 94.4%. This supports the two-level design: group-level control sets how much expert information enters the computation, while within-group control determines which experts receive that budget.

Preference strength. We scan $\gamma \in \{0, 0.5, 1, 1.5, 2\}$ on the same GLUE development samples; $\gamma = 1$ is the pre-specified main configuration. Under a fixed total readout budget, TAME improves from $\gamma = 0$ and remains in a moderate performance range as preference strength increases.

The pre-specified setting lies in this stable region. Increasing γ does not turn the method into hard selection, because all Banks remain available and their local attention masses are still recomputed for each query.

4.4 Generalization to More Challenging Tasks

We next evaluate composition on more challenging mathematics, code, and instruction-following tasks. These domains use different soft-prompt lengths, which TAME can jointly read without padding the experts to a common length.

Table 2: Soft-prompt composition on more challenging tasks. Values are percentages; retention uses the corresponding single context expert as reference.

Method	GSM8K	MBPP	IF	Cross-domain mean	Mean retention
Experts (context)	70.13	38.89	67.47	58.83	—
SP TAME ($\gamma = 1$)	69.90	38.89	56.19	54.99	94.32
SP TAME ($\gamma = 2$)	70.43	38.89	60.07	56.46	**96.49**

The stronger preference improves the balance of retained capabilities across the three domains. Under reference-based group-mass control, increasing γ raises the relative weight of the probe-preferred Bank while keeping all experts available. We interpret this behavior as evidence that preference strength should match the size and specialization of the expert pool: a compact set of specialized experts can benefit from stronger input-level selection, whereas a larger pool containing several relevant experts is better served by softer preferences that preserve broader query-dependent participation. Thus, γ controls how strongly the input-level preference guides the joint output, allowing composition to balance expert selectivity and collaboration.

Together, the GLUE and challenging-task results show that TAME's two-level readout provides a tunable mechanism for preserving and coordinating independently trained soft-prompt experts.

5 Analysis: Cross-Space Comparison and Complementarity

The previous chapter established that TAME improves the joint use of independently trained soft-prompt experts. We now place this result beside parameter-space model merging to characterize two ways of reusing specialized capabilities. The comparison has two purposes: to measure whether contextual composition is competitive at the level of the final model, and to determine whether the two composition spaces preserve the same capabilities. We use a single frozen Qwen3-1.7B backbone and training-free composition throughout, so the comparison isolates where the expert information is stored and how it is read at inference time.

5.1 Comparison with Parameter-Space Merging

Comparison protocol. For the parameter-space route, each task expert is a LoRA update trained on the same frozen backbone. We evaluate Task Arithmetic (TA) (Ilharco et al., 2023), TIES-Merging (Yadav et al., 2023), KnOTS-TIES (Stoica et al., 2025), and TSV-Merge (Gargiulo et al., 2025). For the context-space route, the corresponding expert is an independently trained soft prompt encoded as a KV Bank, and TAME reads the Banks dynamically at each query. Both routes are composed without training a fusion module.

The two expert families have different single-expert operating points, so we report two complementary views. Absolute accuracy measures the final task performance after composition. Capability retention divides the composed score by the corresponding single-expert score; its mean and worst-task value measure how well a method preserves the capabilities it was asked to combine. This paired reporting separates final utility from preservation of the expert portfolio.

Table 3 makes the cross-space result concrete. TAME achieves the highest composed Macro accuracy (77.01%) and the strongest mean and worst-task retention (94.4% and 84.1%) in the comparison. Its Macro accuracy is 1.90 percentage points above LoRA TIES, the strongest parameter-merging baseline by aggregate retention. The result is obtained from soft-prompt experts whose single-expert Macro is 82.19%, whereas the LoRA reference is 83.69%. Thus the contextual route reaches a stronger composed model while starting from a lower expert reference.

Table 3: Cross-space comparison on six GLUE tasks. Each task cell reports accuracy (%) and the change from the corresponding single-expert reference. "Mean ret." and "Worst ret." are the mean and minimum task-level retention.

Method	CoLA	SST-2	QQP	QNLI	RTE	WNLI	Macro	Mean ret.	Worst ret.
Experts (LoRA)	84.08	95.53	91.71	93.85	86.28	50.70	83.69	—	—
Experts (SP)	80.82	95.53	88.90	92.35	84.84	50.70	82.19	—	—
TA	69.61 -17.22	94.15 -1.44	88.24 -3.79	77.28 -17.65	61.01 -29.29	50.70 +0.00	73.50	88.4	70.7
TIES	69.03 -17.90	94.27 -1.32	86.63 -5.55	85.43 -8.97	63.18 -26.78	52.11 +2.78	75.11	90.4	73.2
KnOTS-TIES	70.37 -16.31	93.35 -2.28	89.95 -1.92	81.15 -13.54	57.04 -33.89	54.93 +8.33	74.46	90.1	66.1
TSV-Merge	66.44 -20.98	93.69 -1.92	89.07 -2.89	79.39 -15.41	55.23 -35.98	50.70 +0.00	72.42	87.1	64.0
TAME (ours)	77.66 -3.91	93.46 -2.16	74.76 -15.91	81.04 -12.25	83.03 -2.13	52.11 +2.78	**77.01**	**94.4**	**84.1**

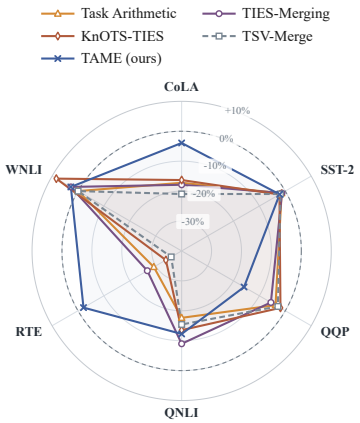

Figure 4: Relative change to own expert reference.

The task columns show why the aggregate result is informative. Parameter merging retains a strong QQP direction, with KnOTS-TIES reaching 89.95%. TAME instead protects the tasks that are most sensitive to shared-parameter interference: it reaches 77.66% on CoLA and 83.03% on RTE, while its worst-task retention remains 84.1%. The two routes therefore make different capability trade-offs, and the contextual route provides the smoother portfolio-level retention profile.

Figure 4 summarizes the retention geometry. The four parameter-space methods produce a common low-retention direction concentrated on CoLA and RTE, despite using different update aggregation rules. TAME does not follow this direction: its reductions are smaller on CoLA and RTE, and its overall polygon is substantially closer to the single-expert reference. This pattern identifies the composition space as a source of the capability-retention profile, rather than a cosmetic difference between merging algorithms.

5.2 Interference and Readout Behavior

The complementary profiles arise from where each route intervenes: parameter merging writes all updates into a shared state, whereas context composition keeps Banks separate and controls readout per query (Fig. 5). The QQP→RTE intervention isolates this difference.

In the parameter route, reducing the QQP task-vector scale from 1.0 to 0.25 raises RTE accuracy from 54.87% to 85.92%, while QQP decreases from 91.11% to 81.15%. A shared update therefore

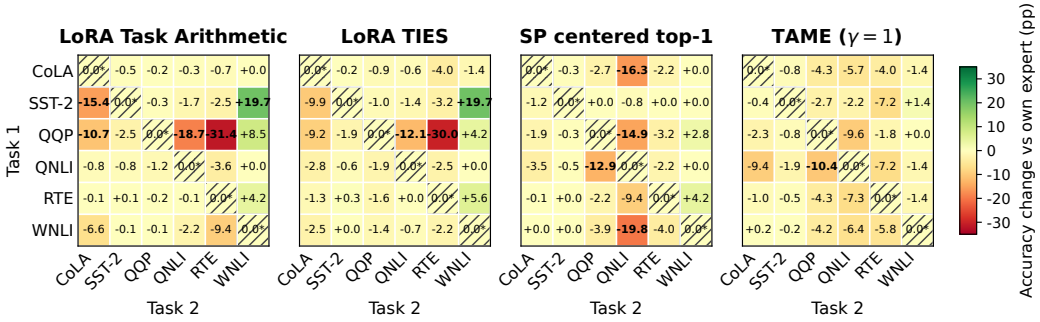

Figure 5: Pairwise composition changes relative to single-expert references. Parameter merging concentrates losses in a few task directions, while controlled contextual readout distributes changes across task pairs.

couples the two capabilities. TAME resolves the same trade-off at readout time: group-level mass control sets how much expert information enters the computation, while within-group allocation assigns that budget to RTE or QQP for the current input.

Mass control recovers the broad degradation of direct joint reading, and preference control changes which Bank receives the fixed budget without rewriting or deleting any Bank. Because the attention masses are recomputed from the current query, different tokens can draw on different mixtures of Banks. This makes parameter interference and contextual readout interference distinct, observable intervention points.

Interpretation. The comparison supports a concrete division of labor. Parameter merging is a compact way to write updates into one model, while soft-prompt composition keeps experts modular and contributions adjustable at inference time. The two routes are parallel, complementary capability-reuse paths with distinct retention geometries and intervention points.

6 Conclusion

We study how to more effectively compose independently trained soft-prompt experts. We propose TAME, a training-free, input-conditioned soft-prompt composition method that separates two sources of imbalance in joint readout: the total amount of expert information entering the model and its allocation across experts. TAME controls these two factors with two levels of attention while preserving the independence of all KV Banks, turning dynamic expert readout into an effective capability-composition mechanism.

Compared with other context-space composition methods, TAME achieves the highest mean capability retention on GLUE among all evaluated baselines. In a unified cross-space comparison, TAME achieves higher GLUE Macro accuracy and mean capability retention than LoRA TIES, the strongest parameter-space merging baseline. On mathematics, code, and instruction-following tasks, TAME also achieves capability retention comparable to or higher than the evaluated parameter-merging methods. Task-level and task-pair analyses further show that the two composition paths preserve different capability directions: parameter merging writes expert updates into a shared parameter state, whereas TAME keeps experts independent and dynamically adjusts their readout at inference time. These results establish TAME-based soft-prompt composition as a competitive training-free capability-reuse path that complements parameter-space model merging.

7 Limitations & Discussions

This work studies parameter-space model merging and context-space soft-prompt composition as two parallel capability-reuse paths, without combining them within a single system. The complementary capability-retention patterns observed in our experiments provide a direct motivation for hybrid designs. For example, parameter merging could write relatively stable general capabilities into the backbone, while independently retained soft-prompt experts could provide input-conditioned specialization through dynamic readout, while preserving a training-free composition procedure.

TAME does not require training an additional fusion module, but contextual composition introduces storage and inference costs. Parameter merging consolidates expert updates into a single model before inference, whereas TAME retains separate KV Banks and performs input-conditioned probing and attention readout at inference time. Bank storage and attention computation therefore grow with the number and length of the Banks. Offline encoding allows the same Banks to be reused across inputs, so practical deployment should consider capability retention together with memory usage and inference latency. Bank compression, lightweight probing, sparse Bank selection, and efficient joint readout are promising directions for scaling soft-prompt composition to larger expert collections.

Our experiments use a unified frozen backbone, a finite collection of experts, and the evaluated task suites, with a focus on composition performance and cross-space capability-retention patterns. Future work can examine larger backbones, larger expert collections, broader task distributions, and adaptive readout strategies under different expert sizes, Bank lengths, and deployment budgets.

Reproducibility Statement

Section 3 defines the Banks and the two-level readout equations, and Section 4.1 introduces the evaluation protocols; Sections 4.2–4.4 specify the soft-prompt composition, ablation, and cross-domain experiments. Section 5 presents the cross-space comparison and its controls. Appendix B documents expert training configurations and the mechanistic analyses, and Appendix A gives the full inference algorithm. These specifications permit reconstruction of the reported accuracy and retention values: the manuscript fixes the calibration protocol, baselines, metrics, and the principal hyperparameters, including the pre-specified preference settings ($\gamma = 1$ on GLUE and $\gamma = 2$ on the cross-domain suite). The release package will include expert training configurations, frozen calibration statistics, evaluation scripts, and run manifests for the reported tables. The raw generations corresponding to the archived sample-level analyses are not included in this writing snapshot.

AI Use Statement

AI tools were used for editorial and typesetting assistance in preparing this manuscript, including conversion to the ICLR format, consolidation of the section files, LaTeX compilation checks, and language and terminology consistency checks. AI tools were not used to generate experimental data, conduct experiments, or replace the authors' scientific judgments. The authors reviewed the manuscript and are responsible for its accuracy, originality, citations, and compliance with the ICLR AI policy.

Ethics Statement

This work studies model merging and context-space expert composition using benchmark tasks and pretrained models. It does not collect new personal data, involve human subjects, or make decisions about individuals. The authors will respect the licenses and usage conditions of the models and benchmarks used in the experiments. We acknowledge that model-composition techniques may be misused, and we report the experimental scope and limitations to support responsible interpretation of the results.

References

Akari Asai, Mohammadreza Salehi, Matthew E. Peters, and Hannaneh Hajishirzi. ATTEMPT: Parameter-Efficient Multi-task Tuning via Attentional Mixtures of Soft Prompts. In *Proceedings of EMNLP*, pp. 6655–6672, 2022. https://arxiv.org/abs/2205.11961.

Robert Belanec, Simon Ostermann, Ivan Srba, and Maria Bielikova. Task Prompt Vectors: Effective Initialization through Multi-Task Soft-Prompt Transfer. *arXiv preprint arXiv:2408.01119*, 2024. https://arxiv.org/abs/2408.01119.

Yuan Cao, Dezhi Ran, Yuzhe Guo, Mengzhou Wu, Simin Chen, Linyi Li, Wei Yang, and Tao Xie. An Empirical Study and Theoretical Explanation on Task-Level Model-Merging Collapse. *arXiv preprint arXiv:2603.09463*, 2026. https://arxiv.org/abs/2603.09463.

Runxi Cheng, Feng Xiong, Yongxian Wei, Wanyun Zhu, and Chun Yuan. Whoever Started the Interference Should End It: Guiding Data-Free Model Merging via Task Vectors. In *Proceedings of ICML*, 2025. https://arxiv.org/abs/2503.08099.

Sabri Eyuboglu, Ryan Ehrlich, Simran Arora, Neel Guha, Dylan Zinsley, Emily Liu, Will Tennien, Atri Rudra, James Zou, Azalia Mirhoseini, and Christopher Ré. Cartridges: Lightweight and General-Purpose Long Context Representations via Self-Study. *arXiv preprint arXiv:2506.06266*, 2025. https://arxiv.org/abs/2506.06266.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task Singular Vectors: Reducing Task Interference in Model Merging. In *Proceedings of CVPR*, 2025. https://arxiv.org/abs/2412.00081.

Vitoria Guardieiro, Avishree Khare, Adam Stein, and Eric Wong. Instruction Following by Principled Boosting Attention of Large Language Models. *arXiv preprint arXiv:2506.13734*, 2025. https://arxiv.org/abs/2506.13734.

Yaru Hao, Yutao Sun, Li Dong, Zhixiong Han, Yuxian Gu, and Furu Wei. Structured Prompting: Scaling In-Context Learning to 1,000 Examples. *arXiv preprint arXiv:2212.06713*, 2022. https://arxiv.org/abs/2212.06713.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing Models with Task Arithmetic. In *International Conference on Learning Representations*, 2023. https://arxiv.org/abs/2212.04089.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of NAACL*, 2019. https://arxiv.org/abs/1902.10186.

Seungeon Lee, Soumi Das, Manish Gupta, and Krishna P. Gummadi. LoRA on the Go: Instance-level Dynamic LoRA Selection and Merging. In *Proceedings of ACL*, 2026. https://arxiv.org/abs/2511.07129.

Brian Lester, Rami Al-Rfou, and Noah Constant. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of EMNLP*, 2021. https://arxiv.org/abs/2104.08691.

Xiang Lisa Li and Percy Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of ACL*, pp. 4588–4599, 2021. https://arxiv.org/abs/2101.00190.

Zongqian Li, Yixuan Su, and Nigel Collier. PT-MoE: An Efficient Finetuning Framework for Integrating Mixture-of-Experts into Prompt Tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. https://arxiv.org/abs/2505.09519.

Zhenyi Lu, Chenghao Fan, Wei Wei, Xiaoye Qu, Dangyang Chen, and Yu Cheng. Twin-Merging: Dynamic Integration of Modular Expertise in Model Merging. In *Advances in Neural Information Processing Systems*, volume 37, 2024. https://arxiv.org/abs/2406.15479.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. No Task Left Behind: Isotropic Model Merging with Common and Task-Specific Subspaces. In *Proceedings of ICML*, 2025. https://arxiv.org/abs/2502.04959.

Michael Matena and Colin Raffel. Merging Models with Fisher-Weighted Averaging. In *Advances in Neural Information Processing Systems*, volume 34, 2021. https://arxiv.org/abs/2111.09832.

Aleksandar Petrov, Philip H.S. Torr, and Adel Bibi. When Do Prompting and Prefix-Tuning Work? A Theory of Capabilities and Limitations. In *International Conference on Learning Representations*, 2024. https://arxiv.org/abs/2310.19698.

George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model Merging with SVD to Tie the Knots. In *International Conference on Learning Representations*, 2025. https://arxiv.org/abs/2410.19735.

Praveen Venkateswaran and Danish Contractor. Spotlight Your Instructions: Instruction-following with Dynamic Attention Steering. *arXiv preprint arXiv:2505.12025*, 2025. https://arxiv.org/abs/2505.12025.

Sarah Wiegreffe and Yuval Pinter. Attention is not not Explanation. In *Proceedings of EMNLP-IJCNLP*, pp. 11–20, 2019. https://arxiv.org/abs/1908.04626.

Zijun Wu, Yongchang Hao, and Lili Mou. Ultra-Low-Dimensional Prompt Tuning via Random Projection. In *Proceedings of EACL*, pp. 1284–1303, 2026. https://aclanthology.org/2026.eacl-long.59/.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving Interference When Merging Models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. https://arxiv.org/abs/2306.01708.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. Tailor: A Soft-Prompt-Based Approach to Attribute-Based Controlled Text Generation. In *Proceedings of ACL*, pp. 410–427, 2023. https://aclanthology.org/2023.acl-long.25/.

Xinyu Yang, Tianqi Chen, and Beidi Chen. APE: Faster and Longer Context-Augmented Generation via Adaptive Parallel Encoding. In *International Conference on Learning Representations*, 2025. https://arxiv.org/abs/2502.05431.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch. In *Proceedings of ICML*, 2024. https://arxiv.org/abs/2311.03099.

Fanhu Zeng, Haiyang Guo, Fei Zhu, Li Shen, and Hao Tang. RobustMerge: Parameter-Efficient Model Merging for MLLMs with Direction Robustness. In *Advances in Neural Information Processing Systems*, volume 38, 2025. https://arxiv.org/abs/2502.17159.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell Your Model Where to Attend: Post-hoc Attention Steering for LLMs. In *International Conference on Learning Representations*, 2024. https://arxiv.org/abs/2311.02262.

Jingang Zhou, Haiyang Guo, Yuan Ma, Han Zhu, and Xu-Yao Zhang. CAMFT: Conflict-Aware Mergeable Fine-Tuning for Large Language Models. *arXiv preprint arXiv:2609.22253*, 2026. https://arxiv.org/abs/2609.22253.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM Knows What You are Looking for Before Generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

A TAME Inference Algorithm

Algorithm 1 summarizes the two-phase procedure of Section 3.3: an uncontrolled probe that fixes $\mathbf{u}(x)$ and the reference Bank $r(x)$, followed by controlled readout in which the per-Bank bias $\beta_k = \gamma u_k + b$ is recomputed at every layer, head, and query while the Banks, $\mathbf{u}(x)$, $r(x)$, and γ remain frozen.

Algorithm 1 TAME Inference for One Input

Require: frozen model f_θ; expert Banks $\{\mathcal{B}_k\}_{k=1}^{T}$; frozen calibration means $\{\mu_k\}_{k=1}^{T}$; preference strength $\gamma \geq 0$; input prompt x
Ensure: generated response y
▷ **Phase 1: input-level probe** (one uncontrolled prefill with all Banks)
$w \leftarrow \min(4,\ |x|)$
$\mathrm{raw}_k \leftarrow$ mean attention mass received by Bank k over the last w queries, all layers and all heads, for all k
$u_k \leftarrow \mathrm{raw}_k - \mu_k$ for all k; $\quad r \leftarrow \arg\max_k u_k$
▷ **Phase 2: controlled readout** (reprocess x, then generate)
for each query position processed under control (reprocessed prompt tokens, then each newly generated token) **do**
  for each layer and each query head **do**
    for each Bank k **do**
      $z_k \leftarrow \log \sum_j \exp(s_{kj})$ ▷ native logits s_{kj} of Bank k
    end for
    $b \leftarrow z_r - \mathrm{logsumexp}_m\big(z_m + \gamma u_m\big)$
    add bias $\beta_k \leftarrow \gamma u_k + b$ to every logit of Bank k ▷ one scalar per Bank
    standard softmax over all visible positions; read the values
  end for
end for
append each new token to the non-expert KV cache; Banks, u, r, and γ stay fixed throughout
return y

B Mechanistic Analysis Details

This appendix provides additional diagnostic analyses related to the composition results: sample-level complementarity (Appendix B.2), the QQP→RTE intervention contrast (Appendix B.3), layer-wise state analysis (Appendix B.4), and generation-time Bank readout trajectories (Appendix B.5).

B.1 Model Configurations and Training Details

All experiments use Qwen3-1.7B as the frozen base model. Parameter-space experts are independently trained with LoRA of rank 16, scaling factor $\alpha = 32$, and dropout 0.05, applied to the query, key, value, output, gate, up, and down projection modules. The same LoRA specification is used across the GLUE and cross-domain settings. Context-space experts are independently trained input-layer soft prompts and are encoded into layer-wise KV Banks before composition; composition itself introduces no additional training.

The original GLUE benchmark contains eight tasks. We retain six tasks for the main comparison because the two excluded tasks did not yield comparably strong SP and LoRA single-task experts under our training setup; using the matched six-task subset keeps the parameter-space and context-space comparisons fair. For the cross-domain experiments, the soft prompts use 256 virtual tokens for mathematics and code, where the tasks are more demanding, and 128 virtual tokens for instruction following. Increasing the prompt length from 128 to 256 improved the mathematics and code experts in our configuration, although their single-expert performance remained below the corresponding LoRA experts. TAME can jointly read Banks with different numbers of virtual tokens, so this domain-specific choice does not require padding the experts to a common prompt length. After composition, the mathematics, code, and instruction-following results still show the retention advantage reported in Table 2.

B.2 Task- and Sample-Level Complementarity

Pairwise setup. We merge the six GLUE experts in all 15 task pairs and evaluate both tasks under each composition method, reporting the accuracy change relative to the corresponding single expert. This isolates the loss caused by composition from expert competence. Centered top-1 hard-routes to the highest-scoring expert from calibrated context-Bank scores, and even this simple operation produces a loss profile qualitatively different from parameter merging.

Sample-level comparison. We compare the post-composition predictions of LoRA TIES and TAME on the GLUE evaluation samples used in Table 1. The samples fall into four categories: both correct, only TIES correct, only TAME correct, and both incorrect. Both uniquely-correct categories are non-empty: each method answers samples that the other misses beyond their shared correct set. Together with the task-level profile in the main table, this is the evidence that the two spaces preserve different capabilities.

B.3 Directional Interference: QQP→RTE

QQP asks whether two questions express the same meaning, whereas RTE asks whether a premise entails a hypothesis; both compare two text spans but require different relational judgments. On the same RTE examples, the single expert is fully correct, TA and TIES nearly lose the task after merging, while TAME preserves most of the correct decisions.

Parameter side: update-strength trade-off. We scale the QQP task vector in the merged parameters while keeping all other settings fixed. Scaling the QQP update from $1\times$ to $0.25\times$ restores RTE accuracy from 54.87% to 85.92%, while QQP accuracy drops from 91.11% to 81.15% on its fixed evaluation set. This identifies the trade-off: writing the QQP donor into shared parameters at a large effective scale changes the comparison direction used by RTE, and the cost of restoring RTE is transferred back to QQP.

Context side: readout strength and expert preference. TAME preserves the independence of the QQP, RTE, and other task Banks while controlling both total Bank mass and within-Bank allocation for every query. Let a denote the total Bank mass, o_T the non-expert text contribution, and $\bar{v}_k$ the conditional readout from Bank k; the readout is

$$o = (1 - a)\, o_T + a \sum_k \rho_k \bar{v}_k, \tag{8}$$

where ρ_k is the within-Bank share for the current query. In QQP→RTE, the mass-only configuration ($\gamma = 0$) already preserves most RTE capability, showing that total-mass control is the first protection against collapse. Biasing the within-Bank preference toward RTE further improves the RTE direction, whereas biasing it toward QQP reintroduces loss, showing that context readout provides a second, input-dependent allocation layer.

The two sides form a clean contrast: adjusting a shared parameter coefficient changes the donor's effective scale everywhere, whereas the two readout controls change only how strongly each independent expert is read at the current query.

B.4 Layer-wise Propagation and State Interventions

Setup. We compare five frozen configurations on the 277 RTE validation inputs of the QQP→RTE pair: the single-task RTE LoRA (86.28%), TA with coefficient 1 (54.87%), TA with the QQP update scaled to $0.25\times$ (85.92%), the single RTE Bank (84.84%), and TAME (82.67%). For each input, we record the hidden state of the last prompt position at the output of each of the 28 Transformer blocks (the final RMSNorm is excluded). Deviations are measured against the reference of each space: the single-task LoRA for parameter-space configurations and the single Bank for context-space configurations. Samples are split by correctness into damaged (reference correct, composed wrong), retained (both correct), repair, and both-wrong groups; curves and interquartile bands are computed within each group.

Parameter side: deviations emerge late. Figure 6 shows the result. Under TA, the damaged and retained groups track each other through the early blocks, separate from block 17, and reach their maximal contrast around blocks 18–22, peaking near block 19 with a $1 - \cos$ deviation of about 0.28 for damaged versus 0.19 for retained samples. Scaling the QQP update to $0.25\times$ suppresses the deviation on the same damaged samples from about 0.28 to about 0.02 across all blocks, restoring RTE to 85.92%. A local check tempers a purely localized reading: scaling the QQP update only in the three-block window $\{17, 18, 19\}$ leaves RTE at 56.32%, close to the 57.04% of a control window $\{14, 15, 16\}$ selected by a frozen rule before seeing accuracies, although the control window carries about 57% larger raw update norms. The deep-block signal is therefore reproducible, but the damage is not repaired by a single localized window, which is consistent with the donor acting through broadly coupled state.

Context side: deviations concentrate early. Figure 7 shows the context-space counterpart, referenced to the single RTE Bank. The contrast between damaged and retained samples is largest in the earliest blocks (0–3), with a secondary peak at block 22. Early divergence is expected in this space: Banks enter the computation from the first block, so expert influence is present before deeper task processing, and the composed readout either steers the trajectory back toward the single-Bank reference or lets it drift.

Interpretation boundaries. The parameter-side damaged group is heavily skewed toward the entailment label (104 of 105 samples), so the block 17–22 signal is reproducible within that subset but is not claimed for both labels. The context-side damaged group is small (n=11) and is reported as a diagnostic rather than a stable mechanism. The curves support the working hypothesis that merged internal representations are coupled to capability damage and weaken when the donor update is scaled; they do not prove that any single block is the sole cause.

Common-state analysis. A common-state analysis separates the perturbation created at the current layer from the state difference inherited from earlier layers: replacing the late-layer state of a merged model with the corresponding state from another configuration and continuing the forward pass isolates the inherited component. After the late-state replacement, the TAME RTE trajectory remains closer to the single-expert trajectory than TIES. The QQP donor therefore enters the shared state and produces task-level consequences downstream, whereas later layers of a parameter merge must continue processing the already mixed direction.

B.5 Dynamic Bank Readout During Generation

A probe-derived preference and anchor do not force the whole answer to read the same Bank. Figure 8 shows the realized Bank readout shares for two correctly answered examples. In the QQP example, QNLI is the anchor, yet the dominant source changes across answer tokens and QQP takes over later. In the RTE example, RTE remains the anchor while the actual shares still change across tokens. Input-level preference and the current query's local matching jointly determine the realized readout, allowing different answer tokens to call different Bank combinations.

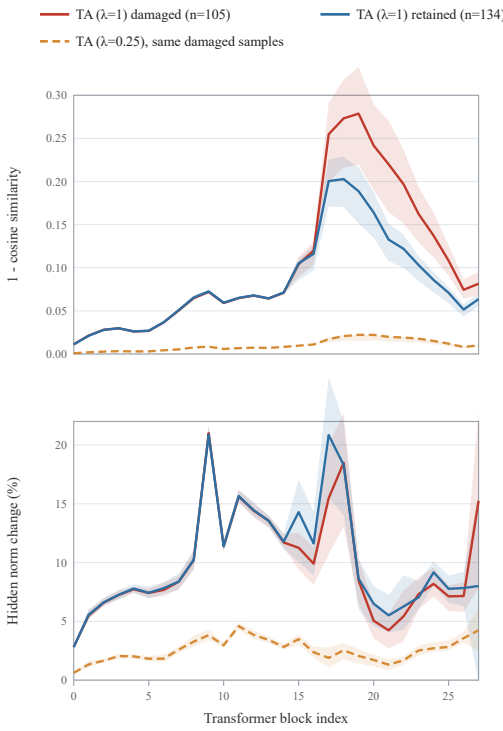

Figure 6: Layer-wise hidden-state analysis on the parameter side (RTE inputs of the QQP→RTE pair, 28 blocks). Top: $1-\cos$ deviation of the last-position hidden state from the single-task RTE LoRA. Bottom: relative change of the hidden-state norm. Curves show group means with interquartile bands for TA samples damaged by merging (n=105) and retained samples (n=134); the dashed curve shows the same damaged samples under TA with the QQP update scaled to $0.25\times$.

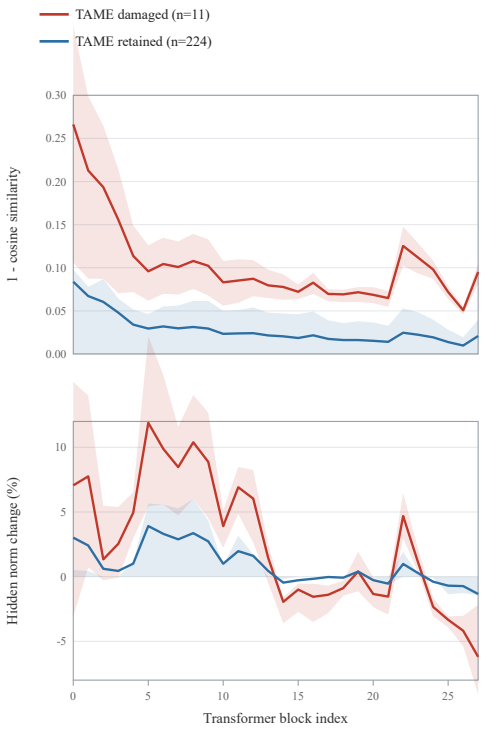

Figure 7: Layer-wise hidden-state analysis on the context side, referenced to the single RTE Bank, with the same inputs and protocol as Fig. 6. Curves show group means with interquartile bands for TAME samples damaged by composition (n=11) and retained samples (n=224). The damage–retention contrast concentrates in the earliest blocks.

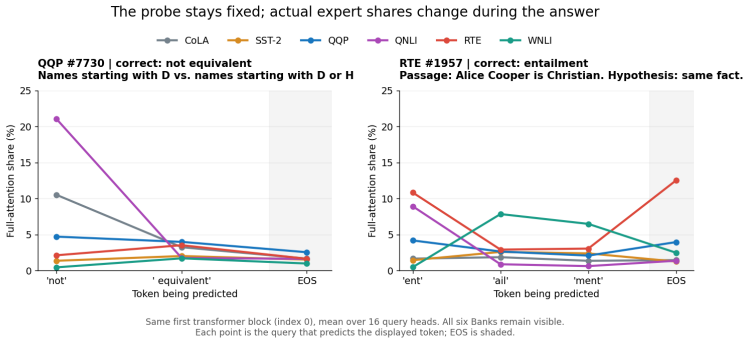

Figure 8: Actual Bank readout trajectories for two correct examples: a QQP input (anchor QNLI) and an RTE input (anchor RTE). Even with a fixed probe and anchor, readout shares change across output tokens; curves show the full-attention share p_k of each Bank.