A TWO-RESOURCE VIEW OF ADAPTER-LIBRARY ARBITRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Growing adapter libraries shift the central challenge of parameter-efficient continual learning from forgetting to arbitration: given an input, which independently trained adapter should answer? Across frozen-language adapter libraries with identical adapters under every arbitration rule, we identify an empirical two-resource pattern. Fitted routers depend on sufficient cross-task meta-training data relative to routing difficulty: holding adapter training fixed, reducing router labels from $1{,}000$ to 100 lowers accuracy from 86.8% to 53.7%; jointly starving adapter training lowers it to 35.4%, while the unfitted competition reaches 83.9% in that cell. Unfitted abstention-based competitions avoid this dependence but pay a different cost: as the library grows, max-over-K arbitration amplifies the upper tail of imperfectly comparable wrong-expert scores. This leads to **Reference-Anchored Slack (RAS) training** for frozen-language libraries: a small shared, domain-adjacent reference trained as slack inside each adapter's existing head, producing more comparable native confidence scores without K-way fitting or refitting. The reference alone adds $+6.3 \pm 3.9$ points to the confidence competition at $K{=}75$ on CLINC150, and its value tracks reference proximity—exchanging an adjacent reference for a distant one costs up to 11.5 points. Within the reference-trained substrate the competition's K-cost stays shallow from 125M to 32B, whereas matched no-reference controls incur larger K-costs at every tested scale, so scale alone does not ensure cross-expert comparability. For generation, the same training principle supplies the shared out-of-scope coordinate that first-token margin arbitration requires. Our results give a regime map for choosing adapter-library arbitration.

1 INTRODUCTION

The unit of continual learning is quietly changing. Whereas classical methods adapt a single network under replay or regularization constraints, an increasingly common pattern—in continual learning (Zhou et al., 2024), multilingual NLP (Pfeiffer et al., 2020), and LLM adapter composition and serving (Huang et al., 2024; Sheng et al., 2024)—is a *library*: a frozen pre-trained backbone plus one parameter-efficient adapter per task, skill, language, domain, or tenant.

Libraries sidestep catastrophic forgetting by construction and make adding capabilities cheap. They concentrate the remaining difficulty into a single question: **given an input, which adapter should answer?**

The prevailing answers fit a router or a combination: an attention-based fusion over adapters (Pfeiffer et al., 2021), a mixture-of-experts gate (Yu et al., 2024), or per-task composition weights fitted on a few examples (Huang et al., 2024). This paper asks when such fitted arbitration works, when an unfitted alternative works instead, and what resource limits each.

The two-resource pattern. Across controlled frozen-language adapter libraries with identical adapters under every arbitration rule (Section 4), two resources govern performance. A *fitted router* depends on its total cross-task meta-training data relative to routing difficulty: with adapters fixed, router-label starvation lowers accuracy $86.8 \rightarrow 53.7\%$ on CLINC150, while on surface-separable intents it stays near 87% to $K{=}75$. An *abstention competition*—each adapter trained with an explicit abstention (*slack*) class, arbitration by max confidence—is the mirror image: it is robust to

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.

meta-data starvation because nothing is fitted across adapters, but degrades as max-over-K arbitration amplifies the upper tail of imperfectly comparable wrong-expert scores. A flat routing oracle confirms that the experts themselves remain capable.

Natural post-hoc repairs do not remove both bottlenecks: the per-adapter calibrations we test do not restore cross-expert comparability, while route-then-verify inherits errors from both routing and verification. These failures motivate *training local discrimination and cross-expert comparability into the same score*.

Reference-Anchored Slack training. We instantiate this principle with **Reference-Anchored Slack (RAS) training** (Section 5.2): a per-expert protocol that labels a small fixed, domain-adjacent reference set as slack inside every adapter's existing head. It requires no cross-adapter fitting and no refitting as the library grows, and the resulting adapters are deployed with their native maximum-class confidence, which the shared reference makes more comparable across experts ($+6.3 \pm 3.9$ points at $K{=}75$, $+2.3 \pm 1.1$ in the starved cell on CLINC150). Controlled within-benchmark swaps show that the gain tracks reference proximity: exchanging only the reference for a distant one costs 11.5 (FewRel) and 4.6 (banking77) points of the confidence competition, and a distant reference is worth no more than no reference at all.

Comparability across scale. Within the reference-trained substrate, larger frozen backbones are associated with thinner wrong-expert score tails and a smaller competition cost, while data-starved fitted routing remains unresolved. Matched three-seed no-reference controls at 1.5B–32B show that scale alone is insufficient: cross-expert comparability depends jointly on the representation and the score-training protocol (Sections 6.4–6.5).

Contributions. (i) **The two-resource pattern.** Under identical-adapter controls on frozen-language libraries (replay budgets 50–1,000, libraries to $K{=}75$), fitted arbitration is limited by its total meta-training data relative to routing difficulty—a fixed-adapter control isolates the router-label effect, and joint buffer starvation is worse still (Table 2)—whereas the unfitted competition is limited by cross-expert score comparability, which the max over K erodes. Three propositions organize the pattern, each term probed by a measurement, and six natural repairs each fail to solve both regimes (Appendix C).

(ii) **Reference-Anchored Slack training.** A no-refit, per-expert training protocol that uses a shared domain-adjacent reference to regularize independently trained experts toward comparable confidence scores. On CLINC150 the reference alone adds $+6.3 \pm 3.9$ points to the confidence competition at $K{=}75$ and $+2.3 \pm 1.1$ in the starved cell; the trained adapters retain old-task accuracy under library growth without any refit, and within-benchmark reference swaps show the gain tracking reference proximity. For generation the same protocol supplies the shared out-of-scope coordinate the first-token margin competition requires.

(iii) **Comparability across scale and a regime map.** Within the reference-trained substrate the competition's K-cost stays inside $[-2.5, -0.1]$ points from 125M to 32B, whereas matched no-reference controls pay -5.1 to -19.4 points: scale alone does not guarantee comparability, while starved fitted routing remains weak at every tested backbone—yielding a regime map for choosing arbitration.

2 Related Work

Adapter libraries and composition. Parameter-efficient adaptation increasingly represents capabilities as separate modules around a shared backbone: AdapterFusion (Pfeiffer et al., 2021) learns attention over task-specific adapters, MAD-X (Pfeiffer et al., 2020) composes language and task adapters, LoraHub (Huang et al., 2024) fits few-shot composition weights, and S-LoRA (Sheng et al., 2024) serves many adapters at scale; EASE (Zhou et al., 2024) and MoE-Adapters (Yu et al., 2024) similarly keep task-specific modules, whereas InfLoRA (Liang & Li, 2024) merges each task's low-rank branch into the backbone. Our focus is orthogonal to how modules are learned: given a fixed library of capable adapters, we study the resources required to select among them at inference.

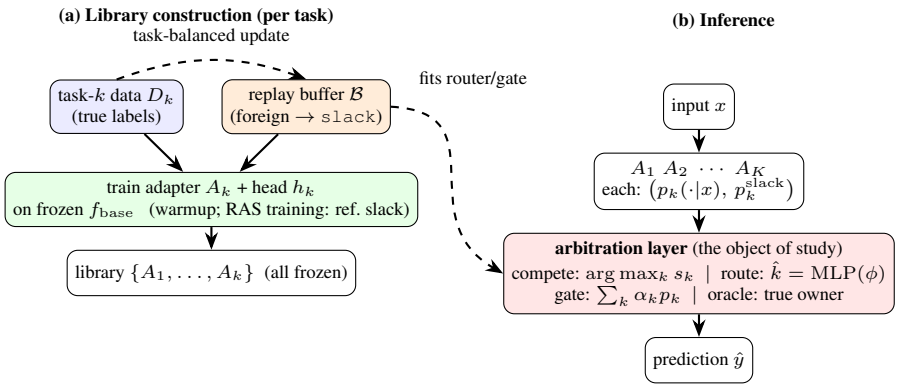


Figure 1: System overview. (a) Each task trains one slack-aware LoRA adapter on a frozen backbone, using buffer samples as foreign (slack) negatives; adapters are never revisited. (b) At inference the adapters produce self-reports and the *arbitration layer* turns them into a prediction; fitted rules train on the same buffer, the competition fits nothing.

Task-free arbitration. Without task identity, a system must resolve which expert owns an input or avoid the choice: EASE (Zhou et al., 2024) matches every adapter's features against class prototypes, SEED (Rypeść et al., 2024) averages all experts' Gaussian class posteriors, and HESTIA (Le et al., 2026) routes by learned density estimates, analysed through density quality and cross-adapter separability. This line is closest to ours; rather than proposing a new selector, we characterize the two resources that limit fitted and unfitted arbitration under identical experts—cross-task meta-training data for fitted routing, cross-expert score comparability for unfitted competition—and introduce a training protocol that repairs the unfitted one.

Abstention and open-set signals. Reject-option classification lets a model abstain rather than force a prediction, from Chow's rule (Chow, 1970) and hinge-loss rejection (Bartlett & Wegkamp, 2008) to open-set recognition (Bendale & Boult, 2016) and none-of-the-above relations (Gao et al., 2019). Outlier exposure (Hendrycks et al., 2019) and energy scores (Liu et al., 2020) offer auxiliary-signal alternatives for single-model OOD detection.

RAS training draws on this tradition but targets a different quantity: its reference is trained as slack not to improve standalone rejection but to give independently trained experts a shared coordinate—*cross-expert comparability* under max-over-K competition.

3 The Adapter-Library Setting

Lifecycle. Figure 1 illustrates the setting. Tasks arrive as a stream. For each task k, a LoRA adapter (Hu et al., 2022) A_k and a small head h_k are trained on the frozen backbone f_{base}. The head predicts the task's classes $\mathcal{C}_k$ plus one explicit abstention output—the *slack* class—in the reject-option tradition (Chow, 1970; Bendale & Boult, 2016). Training uses the task's own labeled examples and foreign samples from a bounded replay buffer, relabeled as slack. The slack class means simply *none of my classes*—not a separate OOD detector, but a local reject option inside each adapter's existing head, suppressing own-class confidence on inputs belonging elsewhere. After training, the adapter and head are frozen, the replay buffer is updated by per-task quota resampling, and the library grows by one; the fitted router and gate consume concatenated logits from all K adapters and so run after adapter evaluation. No frozen adapter or head is ever retrained, so within-expert forgetting is eliminated by construction, though system-level accuracy can still change as arbitration hardens with growth.

Inference. Given an input x, each adapter reports its maximum own-class probability $s_k(x) = \max_{c \in \mathcal{C}_k} p_k(c \mid x)$, which is suppressed when the adapter assigns probability mass to slack. The arbitration layer—the component varied across rules on identical adapters—turns these K self-reports into one prediction. The *raw max-confidence competition* (hereafter *raw competition*) selects the most confident adapter, with no component fitted across adapters; fitted alternatives such as a router

or gate learn a cross-adapter selection function from the replay buffer; and the oracle uses the ground-truth owner.

This isolates the central question: whether independently trained experts can be selected reliably as meta-data, routing difficulty, and library size change. Lifecycle, arbitration procedures, and formal rule definitions (incl. RAS training and its readouts, and the generation margin): Appendices B–D.

4 The Two-Resource Pattern

4.1 Two failure modes, formally

The two failure modes admit simple formal shapes, and each term of the resulting expressions is pinned to a measurement below. Proposition 1 is exact under the stated idealizations (proofs in Appendix A, together with a discussion of the independence assumption); Proposition 2 needs none—both parts are deterministic statements about the transforms; Proposition 3 instantiates a standard multiclass generalization bound.

Proposition 1 (Competition degradation under imperfect comparability)**.** *Fix an owner task k^* and draw X from its test distribution. Let $S^+ = s_{k^*}(X)$ and $S_j^- = s_j(X)$. Assume, as an idealization, that these scores are mutually independent, that the wrong-expert scores share CDF F_-, and that ties have probability zero. The probability of correct arbitration is*

$$P_{\text{comp}}(K) \;=\; \mathbb{E}\big[F_-(S^+)^{K-1}\big],$$

which is non-increasing in K and, for each step $K \geq 2$, strictly decreasing whenever $\Pr\big[0 < F_-(S^+) < 1\big] > 0$ (i.e., unless every input is either perfectly separated or hopeless). Writing $\varepsilon(s) = 1 - F_-(s)$, the probability that no wrong expert exceeds a fixed owner-score threshold s is $(1-\varepsilon(s))^{K-1}$, approximately $e^{-(K-1)\varepsilon(s)}$ for small $\varepsilon(s)$. Thus, small wrong-expert exceedance probabilities can accumulate across competitors.

Proposition 2 (Common-transform invariance and estimated-transform sensitivity)**.** *If the same strictly increasing transform g is applied to every expert score, the arbitration decision is unchanged. For expert-specific transforms g_k and estimates $\widehat{g}_k = g_k + \eta_k$, define the transformed owner–competitor gaps*

$$\Delta_g = g_{k^*}(S^+) - \max_{j \neq k^*} g_j(S_j^-), \qquad \widehat{\Delta}_g = \widehat{g}_{k^*}(S^+) - \max_{j \neq k^*} \widehat{g}_j(S_j^-).$$

Then

$$\big|\widehat{\Delta}_g - \Delta_g\big| \;\leq\; \big|\eta_{k^*}(S^+)\big| + \max_{j \neq k^*} \big|\eta_j(S_j^-)\big|.$$

Proposition 2 separates two cases. A common rank-preserving recalibration cannot change arbitration, whereas expert-specific estimated transforms can perturb the transformed arbitration gap. Empirically, the recalibrations evaluated here do not recover both decision cells at the tested replay budgets: mean shift falls from 83.9 to 52.6% and ECDF to 23.9% under starvation, while they reach 70.3% and 2.4%, respectively, at K=75. Reference-as-slack training instead changes the learned score distributions; the measured wrong-expert margin-score $q_{0.99}$ falls from 5.8 ± 0.4 to 2.7 ± 0.2.

Proposition 3 (Router error at a fixed budget)**.** *A router is a K-way classifier fitted on B samples. For a hypothesis class of Natarajan dimension d, standard multiclass generalization bounds (Shalev-Shwartz & Ben-David, 2014) give excess risk*

$$\mathcal{O}\left(\sqrt{\frac{d\,\log K + \log(1/\delta)}{B}}\right),$$

i.e., the bound decreases as $B^{-1/2}$ for fixed capacity and label count.

The fixed-adapter, fixed-K control isolates router-label availability (Table 2). Banking77 provides the harder-routing comparison: at B=100, the router reaches 46.0% while anchored confidence reaches 74.0% (Section 6.3).

5 METHOD: FROM ABSTENTION TRAINING TO REFERENCE-ANCHORED SLACK

5.1 ABSTENTION TRAINING AND THE RAW COMPETITION

Let a frozen backbone f_{base} accumulate adapters $\{A_1, \ldots, A_K\}$, one per task, each with a head over its own label set $\mathcal{C}_k$ plus one *slack* output. When adapter A_k is trained on task data D_k, we additionally draw foreign samples from the replay buffer and label them slack:

$$\mathcal{L}_k \;=\; \mathbb{E}_{(x,y)\sim D_k}\, \ell\big(h_k(x), y\big) \;+\; \mathbb{E}_{x'\sim \mathcal{B}_{\neg k}}\, \ell\big(h_k(x'), \texttt{slack}\big),$$

so the adapter learns, at training time and from its own perspective, what is *not* its business.

All language classification substrates use a one-epoch learning-rate warmup per adapter.

Given input x, each adapter reports $s_k(x) = \max_{c\in\mathcal{C}_k} p_k(c \mid x)$—its best own-class probability, automatically suppressed by its slack response on foreign inputs—and the prediction is $\arg\max_{c\in\mathcal{C}_{k^*}} p_{k^*}(c \mid x)$ with $k^* = \arg\max_k s_k(x)$. No cross-adapter component is fitted—the *raw competition* baseline. Adding adapter A_{K+1} trains one adapter and touches nothing else, whereas routers must be refitted over $K{+}1$ routes with retained past data (add-without-refit; quantified in the growth scenario of Section 6.2).

5.2 REFERENCE-ANCHORED SLACK: A PER-EXPERT TRAINING PROTOCOL

The slack competition owns local discrimination but lacks a shared score coordinate; post-hoc additions of that coordinate fail, and a separate head sacrifices the discrimination (Appendix C, Tables A1–A2)—so the final method trains both properties into the *same* head and score.

Step 1 (required): reference-as-slack training. Fix once a small, task-independent, *domain-adjacent* reference set R (here: the benchmark's out-of-scope utterances, $|R|{=}250$), shared by every adapter and never changed as the library grows; each adapter keeps its original objective and adds one lightly weighted term teaching the shared reference as slack:

$$\mathcal{L}_k^{\text{RAS}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{foreign}} + \lambda_R\, \mathbb{E}_{x_r\sim R}\, \text{CE}\big(h_k(x_r), \texttt{slack}\big), \qquad \lambda_R = 0.25.$$

No cross-adapter component is fitted and no existing adapter is refitted when the library grows. Hard-negative mining and explicit margin anchoring are separate ablations (Table A6).

Step 2: inference readout. In classification we deploy the RAS-trained adapters with their native maximum own-class confidence $s_k(x)$, selecting $k^*(x) = \arg\max_k s_k(x)$—the raw-competition rule applied to reference-trained adapters, which we call *anchored confidence*—and classify within $\mathcal{C}_{k^*}$. The class–slack *logit margin*

$$e_k(x) = \max_{c\in\mathcal{C}_k} z_{k,c}(x) - z_{k,\texttt{slack}}(x), \qquad k_e^*(x) = \arg\max_k e_k(x),$$

(*anchored margin*) is reported only as a scoring ablation on the identical adapters. In generation no own-class probability is defined, and there the supervised out-of-scope coordinate makes the first-token margin the score (Section 6.5).

What the reference buys. On matched arms (hard-negative mining off on both sides, identical seeds and class-to-task partitions) the shared reference alone lifts the confidence competition by $+6.3 \pm 3.9$ points at $K{=}75$, $B{=}1{,}000$ ($78.2 \pm 4.0 \to 84.5 \pm 0.3$) and by $+2.3 \pm 1.1$ in the starved cell ($83.9 \pm 1.1 \to 86.2 \pm 0.5$), and lowers the measured wrong-expert margin-score tail ($q_{0.99}$: $5.8 \pm 0.4 \to 2.7 \pm 0.2$): under the identical protocol the competition's $K{=}15 \to 75$ cost falls from -6.9 ± 3.2 to -2.5 ± 0.4. Hard-negative emphasis is a separate ablation and contributes none of this (-0.9 ± 1.6 at $K{=}75$, $+0.2 \pm 0.0$ starved). At $\lambda_R = 0.25$, the reference improves the confidence competition in both decision cells.

Readout ablation (classification). Scored on identical adapters, the anchored margin trails the anchored confidence in the starved cell (-2.7 ± 0.6) and shows no consistent advantage at $K{=}75$ (-0.2 ± 1.3; per seed -1.6, $+0.9$, $+0.2$), motivating confidence as the classification default (Table A6, 3 seeds). On plain adapters, whose slack logit is referenced to nothing shared, the same margin is catastrophic (-40.0 ± 8.7 starved, -8.7 ± 4.4 at $K{=}75$): the margin becomes a usable score only once a shared coordinate has been trained—precisely the situation generation forces (Section 6.5).

Table 1: Two scaling axes on CLINC150 (final accuracy %, mean±std over 3 seeds; identical adapters within each run). Budget axis (K=15): fitted rules collapse, raw stays flat. Library axis (B=1000): the fitted family stays flat (total data constant, routing easy) while the competition degrades through cross-expert score-comparability error.

	Budget axis (K=15)			Library axis (B=1000)		
	B=1000	300	100	K=30	50	75
Oracle (ceiling)	99.1±0.0	99.1±0.1	99.1±0.1	99.5±0.1	99.7±0.1	99.8±0.0
Raw competition (unfitted)	85.2±1.0	**85.3**±1.3	**83.9**±1.1	83.0±1.2	80.9±2.5	78.2±4.0
MLP router (fitted)	86.8±0.2	66.2±2.2	35.4±1.9	87.5±0.7	86.8±1.6	87.5±0.5
Soft gate / meta-gate (fitted)	**88.0**±0.2	69.8±2.2	39.2±1.5	**88.2**±0.8	**87.7**±1.5	**88.1**±0.5
Prototype routing	22.9±2.1	20.4±0.6	15.8±1.4	24.5±1.3	29.8±1.3	37.3±0.3

Table 2: Fixed-adapter control (E1): adapters trained once with the full buffer (B=1,000) as negatives; only the router's label pool B_r is starved (CLINC150, K=15; final accuracy %, mean±sample SD over 3 seeds). The competition uses no router labels and is unchanged by construction; the last row is the Table 1 router, whose adapters were starved as well.

Router label budget B_r	1,000	300	100	50
MLP router (adapters fixed)	86.8±0.2	76.1±1.5	53.7±1.5	37.3±2.0
L2 logistic-regression selector	87.1±0.5	79.9±1.2	60.6±2.8	41.1±3.2
Raw competition (fixed adapters)	85.2±1.0	85.2±1.0	85.2±1.0	85.2±1.0
MLP router − raw competition	+1.6	−9.0	−31.5	−47.9
Table 1 router (buffer starves both)	86.8±0.2	66.2±2.2	35.4±1.9	—

6 Results

6.1 Experimental setup

Benchmarks. All main-text benchmarks are language: **CLINC150** (Larson et al., 2019) (150 intents as class-incremental streams, K=15 to K=75; frozen RoBERTa-base, LoRA r8 per task, with the backbone sweep substituting frozen Qwen2.5 decoders), **banking77** (77 mutually similar intents—the hard-routing anchor), **FewRel** (Han et al., 2018) (K=16 relation stream, held-out relations as reference), and **Super-NaturalInstructions** (Wang et al., 2022) (K=20 generation tasks, Section 6.5). Baselines: fitted gate and MLP router, prototype routing, the published systems of Section G, and the oracle ceiling.

Common substrate and rules. Slack training is the shared foundation, not a compared rule: every adapter in a *matched* comparison is slack-trained (published reproductions use their own recipes). Table 1 uses the *plain* substrate (λ_R=0), which defines the *raw competition*; Table 3 crosses reference training ($\lambda_R \in \{0, 0.25\}$) with the confidence and margin readouts, with hard-negative mining disabled, and Table A5 reports the hard-negative-mining ablation over the full grid. The default classification protocol is RAS training without hard negatives (λ_R=0.25), read by *anchored confidence*, with the *anchored margin* as a scoring ablation; the backbone sweep and the growth experiment use this protocol at every scale, and the banking77 and FewRel substrates are specified in Appendix I. Readout comparisons within each training condition use identical adapters; comparisons across training conditions isolate the specified training intervention. Where a table prints plain and anchored rows together (Tables 3 and A5), the row label names the substrate. The rules: the unfitted *raw competition* $k^* = \arg\max_k \max_{c \in \mathcal{C}_k} p_k(c|x)$; fitted *MLP router* (hard selection) and *soft gate* (slack-aware meta-gate fusion), both trained on the buffer over concatenated logits; train-free *prototype routing*; and the ground-truth *oracle* as ceiling (formal definitions in Appendix D).

Protocol. Fixed replay budget (primary) and budget $\propto K$ (ratio control); 3 seeds for all headline tables (single runs marked); identical adapters across rules within each comparison; training-side interventions (RAS training, CR, mining) are reported against the unmodified baseline under the identical protocol, and published-system results (AdapterFusion, O-LoRA) are external validations on their own substrates, not matched comparisons.

Table 3: Reference training × inference readout on CLINC150 (final accuracy %, mean±sample SD over three seeds). Hard-negative mining and the auxiliary competence head are disabled in all four rows. Task partitions are matched across training conditions; the two readouts within each condition use identical adapters. Bold marks the deployed RAS readout.

Training	Readout	K=15, B=100	K=75, B=1000
Plain	Confidence	83.9 ± 1.1	78.2 ± 4.0
Plain	Margin	43.9 ± 8.2	69.5 ± 5.3
RAS training	**Confidence**	$\mathbf{86.2} \pm 0.5$	$\mathbf{84.5} \pm 0.3$
RAS training	Margin	83.5 ± 0.3	84.3 ± 1.1

6.2 Collapse curves at scale

Small-library pilots reproduce the flat oracle and prototype collapse in language (Table A7); Table 1 probes the two scaling axes (Figure A1). On the **budget axis** both fitted rules collapse (router 23.2 at B=50; Figure A1). With adapters fixed, end-to-end accuracy falls from 86.8 to 53.7 at B_r=100 and to 37.3 at B_r=50; at budget 100, jointly reducing the adapter-training buffer yields 35.4, versus 53.7 with adapters held fixed (a *conditional joint-starvation* effect; Table 2)—while the raw competition is flat across B=50–1,000 (83.6–85.3; Table A5); on the **library axis** the router stays flat (total meta-data constant); the soft gate is best at K=75, and what degrades is the competition—the max over 75 imperfectly comparable confidences selects overclaimers (this axis repartitions the fixed 150 intents; with experts and inputs held fixed, growing only the candidate set from 15 to 75 reduces routing accuracy by 2.6±1.5 points within seed (Table A3); the growth experiment below covers true expansion). The pattern is therefore two-dimensional, and the decomposition motivates the method: the competition's weakness is comparability, the results motivate *training* a shared coordinate into each expert; the post-hoc schemes evaluated here did not recover it (Proposition 2).

Isolating reference training from the readout. Table 3 separates reference-as-slack training from the inference readout. With confidence scoring fixed, the reference improves accuracy by 2.33±1.13 points under budget starvation and 6.28±3.88 points at K=75. Replacing confidence with the class–slack margin on the same RAS-trained adapters gives −2.7±0.6 and −0.2±1.3 points, respectively. The effective component in these two cells is reference training, not margin scoring. The broader grid and recipe ablations are reported in Tables A5 and A6.

Library growth: add-without-refit, measured. Growing the library K=50 → 75 makes the system claim concrete (Table A10): anchored confidence loses 1.20 ± 0.09 points of old-task (first-50) accuracy with no refit and the anchored margin 1.00 ± 0.40, while the *refitted* router loses 5.03 ± 0.45 (3 seeds) and ends below the deployed readout (86.6 ± 0.4 vs. 87.5 ± 0.5; $+0.87 \pm 0.17$ within seed)—refitting redistributes fixed capacity over more routes, hurting precisely the tasks it should protect. Add-without-refit is thus an *accuracy-retention* property, not only a cost one.

6.3 The difficulty axis and reference geometry

Banking77 (Casanueva et al., 2020) supplies the hard-routing anchor; both benchmarks here are deployed as Section 5.2 prescribes—RAS training read out by *anchored confidence*. The pattern replicates: at B=100 the fitted rules collapse (router 78.1 → 46.0, gate 79.8 → 49.2 from B=500) and anchored confidence is the only rule left standing (74.0 ± 2.6; 3 seeds), while the *anchored margin* ablation on the identical adapters trails it by 15.7 ± 4.4 points (58.3 ± 4.6)—banking77 carries the CLINC out-of-scope reference, far from its own domain (Figure A2). **RAS training wants a domain-adjacent reference**; FewRel (K=16 relations, held-out relations as the reference) supplies one (Table A11): under starvation anchored confidence leads both fitted rules (57.9 ± 1.4 vs. router 44.2 ± 1.7, gate 47.5 ± 1.2; 3 seeds) and the margin ablation adds $+1.7 \pm 1.1$ on the identical adapters; at B=1,000 the two readouts are within a point (-1.1 ± 0.7) and the fitted rules take the cell (71.8 ± 1.0). Matched arms (reference on or off, one machine per stream) price this directly: FewRel's adjacent reference is worth $+9.9 \pm 1.8$ points of anchored confidence at B=1,000 and $+11.4 \pm 0.8$ at B=100; banking77's far reference, $+0.4 \pm 1.2$ and -1.1 ± 2.6. *Within-benchmark reference swaps* isolate the reference itself (3 seeds; same stream and partitions, no hard-negative mining, only the reference exchanged—it enters training, so the adapters differ).

Table 4: Matched no-reference controls across decoder scales (frozen Qwen2.5, B=1,000, confidence readout; 3 seeds, each seed fixing the identical class-to-task partition in both substrates, hard-negative mining off in both). Δ_K is the K=15 → 75 cost formed within seed; $D_s = \Delta_K^{\text{anch}} - \Delta_K^{\text{plain}}$ is the paired gain of the reference at each scale (percentage points, mean±sample SD). Accuracy levels: Table A4.

Δ_K (confidence scoring)	1.5B	3B	7B	14B	32B
Plain (no reference)	−5.2±2.9	−5.1±2.0	−19.4±2.3	−7.7±0.8	−10.0±1.4
Anchored (RAS-trained)	−0.1±0.6	−0.8±1.0	−1.5±0.7	−1.9±0.3	−1.6±0.9
Paired D_s (mean±std)	5.1±2.6	4.3±1.4	17.8±1.7	5.8±0.6	8.5±0.9

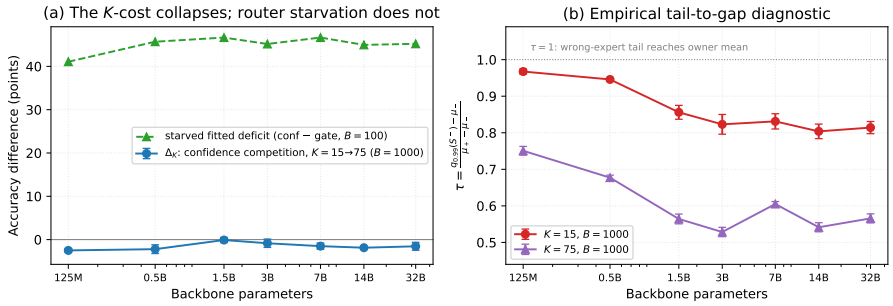


Figure 2: The scale pattern, 125M→32B, on RAS-trained adapters read by anchored confidence (B=1,000 library axis; 3 seeds; error bars are sample SDs of within-seed paired differences in (a) and of per-seed τ in (b)). (a) The K=15 → 75 cost Δ_K is already shallow at 125M (-2.5 ± 0.4) and stays inside $[-2.5, -0.1]$ to 32B, while the starved fitted deficit (anchored confidence − gate, B=100) holds at 41–47 points; the matched no-reference controls (Table 4; RoBERTa plain -6.9 ± 3.2) show that scale alone does not buy the shallow cost. (b) The empirical tail-to-gap diagnostic τ falls across the encoder→decoder transition and then plateaus.

Each stream carries three arms—adjacent, far and no reference—run together on one machine, separating proximity from presence. An adjacent reference lifts anchored confidence by $+5.3 \pm 2.1$ points on a fixed 70-intent banking77 stream (K=10, B=100: $74.4 \pm 2.8 \rightarrow 79.7 \pm 0.8$) and by $+11.0 \pm 1.7$ on FewRel (K=16, B=100: $46.8 \pm 0.5 \rightarrow 57.8 \pm 2.1$), whereas distant references provide no consistent gain over the matched no-reference arms ($+0.7 \pm 2.3$ and -0.5 ± 4.3, respectively). Where such a reference exists (held-out label inventories, OOS logs), RAS training with hard-negative mining puts the confidence competition above its plain counterpart in all seven cells of the CLINC grid and in every seed (Table A5), and above every unfitted alternative in the starved CLINC cell; on FewRel the margin is the stronger readout under starvation (Table A11); absent a reference, the plain competition is the robust default (deployment guidance: Appendix J).

Published systems on the same stream. Against original implementations run on the identical CLINC 15-task stream, anchored confidence leads AdapterFusion as released by 74.7 points in the starved cell and 34.3 at K=75, and O-LoRA as released reaches oracle-grade accuracy only when the test prompt reveals the task, dropping by 36 points once all seen labels compete (Appendix G).

6.4 Comparability across scale

We sweep the backbone—RoBERTa-base (Liu et al., 2019) (125M) and frozen Qwen2.5 (Yang et al., 2024) decoders, 0.5B to 32B (bf16, one LoRA per task; largest on a single H200)—running both decision cells plus a fixed-budget K=15, B=1,000 control at every scale (Figure 2; Tables A9–A12).

Two regularities separate. First, **the pattern persists across the tested backbones**—starved fitted rules collapse (router 21.8–39.8, gate 34.7–47.6 vs. anchored confidence 80.4–92.8; deficit 41–47 throughout) and recover with budget (Table A9: B:100 → 1,000 restores the router by 46–55; 3 seeds)—scaling alone does not rescue the fitted routers under the evaluated budget and protocol.

Second, **on RAS-trained adapters the K-cost is shallow at every tested backbone**: at fixed B=1,000 the anchored-confidence K=15 $\rightarrow$ 75 cost is -2.5 ± 0.4 at 125M, -2.2 at 0.5B, then floats in $[-1.9, -0.1]$ to 32B—a property of the substrate rather than a collapse bought by scale. Matched no-reference controls pay a larger K-cost at every scale (-5.2 to -19.4 vs. -0.1 to -1.9, 1.5B–32B, Table 4; RoBERTa plain -6.9 ± 3.2 against -2.5 ± 0.4 on identical partitions)—the substrate, not margin scoring, carries this robustness.

Figure 2(b) reports an empirical tail-to-gap diagnostic motivated by Proposition 1, $\tau = \big(q_{0.99}(S^-) - \mu_-\big)/\big(\mu_+ - \mu_-\big)$, measured in the margin coordinate. At 125M $\tau = 0.97 \pm 0.01$; across the encoder$\rightarrow$decoder transition it falls to 0.82 and plateaus, tracking the residual K-cost: consistent with the tail mechanism. The anchored margin adds nothing at any measured scale—it trails anchored confidence by 4.1 ± 0.3 at 125M and by ≤ 1.6 points at 1.5B and 7B–32B—which is why classification deploys the confidence readout; in generation the coordinate must first be constructed before any margin exists.

6.5 Generation: RAS training makes the first-token margin well-defined

Libraries increasingly serve *generation*. We build one: K=20 Super-NaturalInstructions tasks (Wang et al., 2022), one LoRA per task on frozen Qwen2.5-1.5B, winner-only greedy decoding, Rouge-L F1; the reference-as-slack principle is instantiated for generation—foreign replay and a 100-input reference drawn from five held-out tasks of the same instruction corpus enter each expert's training pool with a designated out-of-scope literal as their target, so every expert is supervised to emit it—and the first-token logit margin against that literal is the score.

The first finding is structural: **without the reference, or another mechanism that supervises every expert's slack coordinate, the first-token-margin competition is broken by construction**—the first expert's slack coordinate is never positively supervised, and routing is near chance regardless of budget ($7.5 \pm 3.8\%$; chance 5.0)—an asymmetry classification hid but generation exposes. With RAS training the identical rule routes at $85.3 \pm 7.4\%$ (87.6 on the matched draw; Table A13); a guard-only control (first expert only) recovers only partially—this first-token competition requires the shared coordinate under the evaluated lifecycle (Appendix C.4). This experiment supports shared abstention supervision for first-token arbitration. Unlike the classification starvation cells, SNI remains easy to route at B=100: the router reaches 88.7% routing accuracy versus 85.3% for the anchored competition, while both attain mean system Rouge-L of 60.2 (Table A13; Appendix C.4).

Scope: vision stress tests. Appendix H reports two vision stress tests on CIFAR-100—a frozen-backbone adapter library and a separately implemented trainable full-network system with no adapters—and neither shows transfer of RAS training: on the adapter library the reference does not move the confidence competition (-0.29 ± 0.47 against the matched no-reference arm, paired over 3 seeds) while the anchored margin collapses (17.36 ± 0.42, task-incremental ceiling intact at 83.00 ± 0.08), and on trainable representations the reference term destroys first-task accuracy (-30.67 ± 14.18). These results delimit the demonstrated positive scope to frozen-language libraries.

7 Conclusion

Growing adapter libraries make arbitration a central problem of parameter-efficient continual learning. Our experiments identify two bottlenecks—meta-training data for fitted routers and cross-expert comparability for unfitted competition—and show that domain-adjacent reference-as-slack training improves confidence arbitration in frozen-language libraries without K-way fitting or refitting. The resulting regime map is practical: deploy a fitted router where cross-task meta-data is plentiful and routing is easy, and the anchored confidence competition where meta-data is scarce, routing is hard, or the library must grow without refitting. Extending reference-as-slack training to representations that are still being trained, and finding domain-adjacent references for vision libraries, are the natural next steps (limitations: Appendix F).

Reproducibility Statement

All experiments run on a single GPU per run (a single NVIDIA H200 for the largest backbones), entirely from public assets: CLINC150, banking77, FewRel, Super-NaturalInstructions, CIFAR-100, SVHN (the vision reference set); frozen RoBERTa-base and Qwen2.5 (0.5B–32B) backbones; published original code for AdapterFusion (`adapters` library (Poth et al., 2023)) and O-LoRA. Every table states its seed count; single-run entries are marked. Training recipes (LoRA r=8, α=16 on attention query/value projections for all language backbones, λ_R=0.25, buffer protocol) are specified in Sections 3–5 and Appendix I, and the language-side experiment scripts (library construction, the arbitration rules of Appendix D, the reference-swap and guard controls, and the table, figure and number-check scripts, with a per-table command map) are included as supplementary material together with the run outputs behind every language number, so each language table and figure regenerates from the archive and a verification script re-derives every printed cell. The code for the two vision testbeds, with their extractors and code-freeze manifests, will be released upon publication.

AI Use Statement

We used generative AI tools, under the authors' direction, for the following tasks with required disclosure: feedback on hypotheses, methodology, and control experiments; method implementation; dataset reformatting; proof-writing assistance; and interpretation of results. We did not use them to generate synthetic data, formulate the central mathematical claims, translate, or perform qualitative analysis. We also used them for recommended-disclosure tasks (drafting and editing, literature search, reference formatting, figures, code review). All AI-assisted work was verified by the authors—code and analyses inspected, proofs checked, and all reported numbers produced by our own experimental pipeline—and we take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9:1823–1840, 2008.

Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1563–1572, 2016.

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

Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. New insights on reducing abrupt representation change in online continual learning. In *International Conference on Learning Representations*, 2022.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 38–45, 2020.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. FewRel 2.0: Towards more challenging few-shot relation classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6250–6255, 2019.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4803–4809, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. LoraHub: Efficient cross-task generalization via dynamic LoRA composition. In *First Conference on Language Modeling (COLM)*, 2024.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1311–1316, 2019.

Hang Thi-Thuy Le, Nam-Quan Nguyen, Lam-Huy Nguyen, Dien Dinh, Minh Hoang, and Trong Nghia Hoang. Task-free continual learning via order-invariant linearized adaptation and density-guided adapter routing. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 3274–3297. PMLR, 2026.

Yan-Shuo Liang and Wu-Jun Li. InfLoRA: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23638–23647, 2024.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7654–7673, 2020.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 487–503, 2021.

Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. Adapters: A unified library for parameter-efficient and modular transfer learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 149–160, 2023.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5533–5542, 2017.

Grzegorz Rypeść, Sebastian Cygert, Valeriya Khan, Tomasz Trzciński, Bartosz Zieliński, and Bartłomiej Twardowski. Divide and not forget: Ensemble of selectively trained experts in continual learning. In *International Conference on Learning Representations*, 2024.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Ying Sheng, Shiyi Cao, Dacheng Li, Coleman Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, Joseph E. Gonzalez, and Ion Stoica. S-LoRA: Serving thousands of concurrent LoRA adapters. In *Proceedings of Machine Learning and Systems*, volume 6, pp. 296–311, 2024.

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, Singapore, 2023.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, et al. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5085–5109, 2022.

An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23219–23230, 2024.

Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23554–23564, 2024.

A PROOFS AND ASSUMPTIONS

Setting and assumptions. Fix an owner task k^* and draw X from its test distribution. For a frozen library, define $S^+ = s_{k^*}(X)$ and $S_j^- = s_j(X)$. Assume (A1) mutual independence of these scores over the input draw, (A2) a common wrong-expert CDF F_-, and (A3) continuity, so ties have probability zero. A1 is an idealization: the scores share the same query and backbone and need not be independent in practice. The proposition describes the mechanism—degradation driven by the wrong-expert upper tail—which our diagnostics measure directly (the wrong-expert 99th percentile falls $5.8 \pm 0.4 \to 2.7 \pm 0.2$ under reference-as-slack training, and the corresponding arbitration gain follows). For heterogeneous independent wrong experts, the product becomes $\prod_j F_{-,j}(S^+)$. Adding competitors remains non-improving when the existing score distributions are held fixed; strictness depends on the added experts.

Proof of Proposition 1. By A1–A3,

$$P_{\text{comp}}(K) = \Pr\left[S^+ > \max_{j \le K-1} S_j^-\right] = \mathbb{E}_{S^+}\left[\prod_j \Pr[S_j^- < S^+ \mid S^+]\right] = \mathbb{E}\big[F_-(S^+)^{K-1}\big].$$

Since $0 \le F_-(S^+) \le 1$ almost surely, the integrand is non-increasing in K pointwise, so $P_{\text{comp}}(K+1) \le P_{\text{comp}}(K)$. Strict decrease fails exactly when $F_-(S^+) \in \{0,1\}$ almost surely—every input either perfectly separated ($F_-(S^+) = 1$) or hopeless ($F_-(S^+) = 0$)—and holds whenever $\Pr[0 < F_-(S^+) < 1] > 0$. For the tail form, write $\varepsilon(s) = 1 - F_-(s)$; then $F_-(s)^{K-1} = e^{(K-1)\log(1-\varepsilon(s))}$ and, for $\varepsilon(s) \le \frac{1}{2}$,

$$e^{-(K-1)\varepsilon(s)(1+\varepsilon(s))} \;\le\; F_-(s)^{K-1} \;\le\; e^{-(K-1)\varepsilon(s)},$$

or equivalently,

$$e^{-(K-1)\varepsilon(s)^2} \;\le\; \frac{F_-(s)^{K-1}}{e^{-(K-1)\varepsilon(s)}} \;\le\; 1.$$

The exponential approximation is pointwise in the owner-score threshold. Reducing ε—a property of the trained score distributions—raises the integrand at every threshold. □

Proof of Proposition 2. If the same strictly increasing function g is applied to every score, then

$$g(S^+) > \max_j g(S_j^-) \iff S^+ > \max_j S_j^-,$$

so the winning expert and P_{comp} are unchanged. For the expert-specific case, let

$$a = g_{k^*}(S^+), \qquad \widehat{a} = a + \eta_{k^*}(S^+),$$

and

$$b = \max_{j \ne k^*} g_j(S_j^-), \qquad \widehat{b} = \max_{j \ne k^*}\big(g_j(S_j^-) + \eta_j(S_j^-)\big).$$

The max operator is Lipschitz under the sup norm:

$$|\widehat{b} - b| \;\le\; \max_{j \ne k^*} |\eta_j(S_j^-)|.$$

Therefore,

$$\begin{aligned} |\widehat{\Delta}_g - \Delta_g| &= |(\widehat{a} - \widehat{b}) - (a - b)| \\ &\le |\widehat{a} - a| + |\widehat{b} - b| \\ &\le |\eta_{k^*}(S^+)| + \max_{j \ne k^*} |\eta_j(S_j^-)|. \end{aligned}$$

□

Proposition 3: provenance. The bound is the standard agnostic-PAC guarantee for multiclass ERM over a hypothesis class of Natarajan dimension d (Shalev-Shwartz & Ben-David, 2014, Ch. 29), instantiated with routes as labels and the replay buffer (size B) as the meta-training sample. We use the bound to motivate sample-budget sensitivity at fixed capacity and label count, not to predict the measured error rate. The causal budget evidence comes from the fixed-adapter, fixed-K experiment. Routing difficulty is assessed empirically through the benchmark comparisons in Section 6.3.

B LIFECYCLE AND ARBITRATION ALGORITHMS

Algorithm 1 Library construction with slack training

Require: frozen backbone f_{base}, buffer capacity B, shared reference set R, weight λ_R

$\mathcal{B} \leftarrow \emptyset$
for task $k = 1, 2, \ldots$ **do**
  $S_k \leftarrow$ samples from $\mathcal{B}$, relabeled `slack`
  train (A_k, h_k) on $D_k \cup S_k$ over frozen f_{base}, adding $\lambda_R \, \mathbb{E}_{x_r \sim R} \, \text{CE}(h_k(x_r), \texttt{slack})$ {1-epoch LR warmup; λ_R=0 for plain, 0.25 for RAS classification}
  freeze (A_k, h_k); add to library
  if $k = 2$ **and** guard enabled {off in every language classification run; generation control and full-network vision testbed only} **then**
    warm-start guard retrain of A_1 {re-optimize A_1 from its own weights, lr×0.1, on $\mathcal{B}$ (true labels) + D_2 (slack)}
  end if
  $q \leftarrow \lfloor B/k \rfloor$
  retain up to q uniformly sampled items from each previous task in $\mathcal{B}$ and add up to q uniformly sampled items from D_k {task-balanced quota; the realized buffer can be smaller than B}
end for

Algorithm 2 Arbitration at inference (rule ∈ {compete, margin, route, gate, proto, oracle}). On RAS-trained adapters, compete is the *anchored confidence* readout—the classification default—and margin the *anchored margin* ablation.

Require: input x, library $\{(A_k, h_k)\}_{k=1}^{K}$

for $k = 1$ **to** K **do**
  $\mathbf{z}_k \leftarrow h_k\big((f_{\text{base}} + A_k)(x)\big)$; $p_k \leftarrow \text{softmax}(\mathbf{z}_k)$; $s_k \leftarrow \max_{c \in \mathcal{C}_k} p_k(c)$
end for
$\phi(x) \leftarrow [\mathbf{z}_1, \ldots, \mathbf{z}_K]$ {concatenated logits; the fitted rules consume all K adapters}
if rule = compete **then**
  $k^* \leftarrow \arg\max_k s_k$ {unfitted; slack suppresses foreign s_k}
else if rule = margin **then**
  $k^* \leftarrow \arg\max_k \big(\max_{c \in \mathcal{C}_k} z_{k,c} - z_{k,\texttt{slack}}\big)$ {class–slack logit margin; *anchored margin* on RAS-trained adapters (Section 5.2)}
else if rule = route **then**
  $k^* \leftarrow \text{MLP}_{\text{route}}\big(\phi(x)\big)$ {fitted on buffer, owner labels}
else if rule = gate **then**
  $\boldsymbol{\alpha} \leftarrow G\big(\phi(x)\big)$;
  return $\arg\max_c \sum_k \alpha_k \, p_k(c)$ {soft fusion (meta-gate)}
else if rule = proto **then**
  $k^* \leftarrow \arg\max_k \cos\big(e(x), \mu_k\big)$ {$e(x)$: frozen-backbone embedding}
else if rule = oracle **then**
  $k^* \leftarrow$ true owner of x
end if
return $\arg\max_{c \in \mathcal{C}_{k^*}} p_{k^*}(c)$

C ENGINEERED COMPOSITES: ROUTE-THEN-VERIFY AND COMMON-REFERENCE ARBITRATION

C.1 POST-HOC RECALIBRATION: DEFINITIONS AND VERDICTS

Four self-calibration repairs re-express each expert's competition score *after* training; together with Route-then-Verify below they are the five post-hoc failures referenced in the main text. *Mean-shift* subtracts each expert's buffer-estimated mean score ($83.9 \rightarrow 52.6\%$ starved; $78.2 \rightarrow 70.3\%$ at $K{=}75$). *Per-expert ECDF* replaces the score by its rank within expert-specific calibration samples ($\rightarrow 23.9\%$ starved; $\rightarrow 2.4\%$ at $K{=}75$). *Per-expert temperature* rescales each expert's logits by a buffer-fitted temperature (no rescue in language: $\rightarrow 75.1 \pm 5.5$ starved and $\rightarrow 70.6 \pm 4.8$ at $K{=}75$; see Appendix H for the vision adapter-testbed counterpart). The *common-protocol conformal transform* re-expresses each score as a tail probability over the shared reference (a no-op; Section C.3). Proposition 2 shows that common rank-preserving recalibration cannot change arbitration and that estimation errors in expert-specific transforms perturb the transformed score gap. Empirically, the expert-specific recalibration methods evaluated here do not repair both decision cells at the tested replay budgets.

C.2 ROUTE-THEN-VERIFY: COMPOSING THE TWO INVARIANCES

Running all K adapters costs $O(K)$ forwards; a cheap top-m pre-filter can bound inference, requiring only recall@m rather than rank-1 accuracy—conditional on an independently verified shortlist source, which neither pre-filter tested below satisfied. The two-resource pattern of Section 4 exposes complementary failure modes: a fitted router fails by *wild misrouting* when its total meta-data is starved, while the abstention competition fails through the *max over K imperfectly comparable confidences* as the library grows. Each mechanism is invariant to exactly the resource that destroys the other—which suggests composing them so that each is used only where it is robust.

Route-then-Verify (RTV) arbitrates in two stages: a *shortlist* source proposes the top-m candidate adapters, and the slack-mediated abstention competition decides *within* the shortlist. The division of labor is deliberately asymmetric. The shortlist is only required to place the true owner in its top m—a recall@m requirement that is far weaker than rank-1 routing accuracy, and that even the otherwise-collapsing prototype signal may satisfy—while the competition, restricted to $m \ll K$ candidates, pays the calibration-noise cost of an m-way rather than K-way max, returning it to the small-library regime where it is near-ceiling. When the shortlist source is the prototype similarity (a per-task centroid, updated incrementally), the composed system retains the add-without-refit property in full: adding an adapter adds one centroid and touches nothing else. Choosing m trades the two error sources (shortlist misses vs. residual comparability noise); inference cost is $O(m)$ adapter passes plus one cheap embedding for the prototype shortlist (the MLP shortlist instead consumes all K adapters' logits, so it bounds nothing).

Verdict: sequential composition inherits the minimum, not the maximum. The decision experiments refute RTV in both target cells, and the reason is instructive (Table A2). In the large-library cell ($K{=}75$), the best RTV configuration reaches 82.3%—above the pure competition (78.2%) but below the pure router (87.5%): with the shortlist already 97% accurate, the m-way verification step subtracts more precision than it adds. In the budget-starved cell ($K{=}15$, $B{=}100$), RTV collapses to 70.2% against the pure competition's 83.9%: the starved router's *entire ranking* is corrupted, not merely its top-1 (recall@10 is only 81%), so using it as a hard pre-filter excludes the true owner before the competition can recover it. A sequential pipeline thus inherits the *worse* of (shortlist recall, verifier precision) in every regime—the composition binds the failure modes together instead of trading them away. Together with the measured self-calibration failures (mean-shift $83.9 \rightarrow 52.6$, ECDF $\rightarrow 23.9\%$ in the starved cell; $78.2 \rightarrow 70.3/2.4\%$ at $K{=}75$), across the tested variants, neither recalibrating the competition nor filtering it through a router escaped the two-resource pattern. The constructive conclusion of this paper is therefore *regime-matched arbitration*: deploy the unfitted abstention competition where meta-data is scarce, routing is hard, or refit-free growth is required, and a fitted router where total meta-data is adequate and routing is easy—with the deployment guidance of Appendix J as the decision instrument.

C.3 COMMON-REFERENCE ARBITRATION: TRAINING COMPARABILITY IN

The repair attempts above share a flaw the pattern itself predicts: they operate *after* training, on scores whose coordinate systems were never aligned. The two-resource decomposition suggests a design principle instead:

> *A scalable arbitration mechanism should require no K-way global fitting, while producing expert scores that are comparable by construction.*

Common-Reference Arbitration (CR-SLoRA) realizes this by giving every adapter, alongside its class head, a scalar competence head g_k trained to discriminate its own task data from one *fixed reference distribution* q shared by all adapters (in our language testbed, the CLINC out-of-scope utterances, $|R|$=250):

$$\mathcal{L}_k = \mathcal{L}_{\text{cls}} + \lambda\big[\text{BCE}(g_k(x_{\text{native}}), 1) + \text{BCE}(g_k(x_r), 0)\big], \qquad k^*(x) = \arg\max_k g_k(x).$$

Under optimal discrimination—reading g_k as the pre-sigmoid logit, with balanced native/reference sampling so the constant C is common across adapters—$g_k^*(x) = \log p_k(x)/q(x) + C$, and the shared denominator makes $\arg\max_k g_k^* = \arg\max_k p_k$: comparability holds by construction, no K-way component is fitted, and because q never changes, adding adapter $K{+}1$ leaves every existing g_k's target invariant—add-without-refit as a proposition rather than a property claim.

The two decision cells split the verdict (Table A1). In the large-library cell, CR-SLoRA recovers $+7.0 \pm 2.4$ points over the same-run confidence score ($75.7 \to 82.7$ at K=75; $84.0 \to 89.2$ at K=40), with the score-gap diagnostic improving in lockstep—the first substantive repair of the competition's K-cost after five post-hoc failures, and direct evidence for the principle that comparability must be *trained in*. In the budget-starved cell it fails (67.7 vs. 79.5), and the diagnosis is again instructive: the reference q lies far from every task, so g_k learns an "in-domain vs. out-of-scope" boundary and is uncontrolled exactly on foreign in-domain inputs—an extrapolation region between the two training distributions; with only $\sim$100 buffer negatives the hybrid variant cannot close the hole. The competition's raw confidence, which is strong precisely in that cell, coexists in the same system at no extra cost, making a fixed, fitting-free evidence combination $r_k = g_k + \log s_k$ a natural candidate—but the tested additive combination fails the same way: it is harmless where g is healthy (83.1 vs. 82.7 at K=75) and inherits g's corruption where it is not (70.8 vs. raw 79.5 in the starved cell). The remaining principled option is the common-protocol conformal transform: at inference each adapter re-expresses its score as a tail probability within its own distribution over the *shared* reference R, $u_k(x) = \hat{P}_{r\sim R}[g_k(r) < g_k(x)]$—unlike the failed per-expert ECDF, whose reference sets differed across experts, the calibration protocol here is common to all. It changes nothing in either cell (67.7 and 82.4, both within noise of the untransformed score), and the reason completes the diagnosis: the starved-cell failure is a *value* problem, not a *scale* problem—g_k ranks foreign in-domain inputs incorrectly *within* each expert, so the per-expert monotone transforms evaluated here did not repair the cross-expert ordering at replay-buffer scale. The starved cell's deficit is in-domain negative *coverage*, which is precisely the data that the budget constraint withholds; comparability is trainable, but the scarcity was not repaired by the scoring transformations tested here. The separate-head variant's final position is therefore at best a *partial* repair: it recovers part of the large-library cost ($+4.5 \pm 4.6$, negative in one of three seeds) but loses the local discrimination that the slack head already possessed. Jointly training the separate competence head also depresses the confidence competition on the reference-trained substrate by about six points in both decision cells (-6.3 starved, -6.7 at K=75; on the plain substrate -4.3 and -2.5), even though that head is never used for confidence inference—one reason to train the shared coordinate inside the existing head.

C.4 GENERATION WELL-POSEDNESS: GUARD CONTROL

Classification hid this asymmetry (own-class softmax probabilities stay informative with an untrained slack unit); generation exposes it—the first-token distribution *is* the score—and the guard-only control (the classification lifecycle's warm-start guard applied to the first expert; 3 seeds, each a fresh task draw) measures what repairing one expert buys: routing recovers from chance only partially and unstably ($77.1/30.8/74.6\%$, consistent with the asymmetry surviving in other under-exposed experts), whereas RAS training wins on every matched draw ($87.6/91.3/77.1\%$). With RAS training the identical rule routes at $85.3 \pm 7.4\%$ over 3 seeds (87.6% on the matched draw,

Table A1: Common-Reference Arbitration in the two decision cells (CLINC150, final accuracy %, mean±sample SD over 3 seeds; "pure" trains g_k against the reference only, "hybrid" adds in-batch buffer-foreign negatives; both variants train on the plain substrate, λ_R=0, so g_k is the only reference-aware component). The confidence row gives the range of the two runs' own confidence competitions. Score-gap diagnostics: fraction of inputs with $S^+ > \max S^-$.

	K=75, B=1000	K=15, B=100
Confidence score (same run)	75.7–77.7	79.5–80.6
CR-SLoRA (pure)	**83.1**±1.3	63.2±3.3
CR-SLoRA (hybrid negatives)	82.7±0.6	67.7±0.4
Soft gate (fitted, same run)	86.9±0.9	46.4±1.9
$S^+ > \max S^-$ fraction: raw → CR	75.7 → 82.7	79.6 → 68.2

Table A2: Route-then-Verify decision experiments (CLINC150, plain substrate, final accuracy %, mean±sample SD over 3 seeds; recall@m of the selected shortlist in parentheses). Sequential composition loses to the better pure mechanism in both cells.

	K=75, B=1000	K=15, B=100
Pure router (MLP)	**87.5**±0.5	35.4±1.9
Pure abstention competition	78.2±4.0	**83.9**±1.1
RTV (MLP shortlist, best m)	82.3±2.0 (97.4)	70.2±2.3 (80.9)
RTV (prototype shortlist, best m)	64.1±1.1 (72.7)	66.4±2.1 (75.7)

Table A13). The first expert receives no positive slack-target supervision without the reference; the shared reference supplies this supervision to every expert and outperforms the first-expert-only guard on all three matched task draws. In generation, consistently supervising the abstention coordinate is required for the proposed first-token margin under this lifecycle. SNI remains easy to route with five buffer samples per task: at B=100, the router and the anchored competition have equal reported mean Rouge-L (60.2), while the router has higher mean routing accuracy. A confidence-only competition without the slack signal trails both everywhere, so the trained slack coordinate, not generic confidence, is what carries the comparability here.

D Arbitration Rules: Formal Definitions

Arbitration rules, formally. Let each adapter A_k output a probability vector $p_k(\cdot \mid x)$ over its own classes $\mathcal{C}_k$ plus the slack output $p_k^{\text{slack}}(x)$. The rules under study are:

- *Raw max-confidence competition* (unfitted; the *raw competition* of the main text): $s_k(x) = \max_{c \in \mathcal{C}_k} p_k(c \mid x)$; predict with $k^* = \arg\max_k s_k(x)$.
- *MLP router* (fitted, hard): features $\phi(x) = [\mathbf{z}_1(x), \ldots, \mathbf{z}_K(x)]$ (concatenated logits, slack logits included); an MLP trained on buffer samples with owner labels predicts $\hat{k}(x)$, which then predicts within $\mathcal{C}_{\hat{k}}$.
- *Soft gate* (fitted, soft): a gate G produces a distribution over adapters, $\boldsymbol{\alpha}(x) = G(\phi(x))$ with $\sum_k \alpha_k(x) = 1$, and the prediction is the weighted fusion $\hat{y} = \arg\max_c \sum_k \alpha_k(x)\, p_k(c \mid x)$. In the full meta-gate form, ϕ additionally contains slack-aware statistics per adapter (slack probability, maximum and sum of non-slack probabilities, entropy, top-1/top-2 margin), and G is trained by KL divergence to soft targets $w_k(x, y) \propto p_k(y \mid x)\big(1 - p_k^{\text{slack}}(x)\big)$, with the convention $p_k(y \mid x) := 0$ whenever $y \notin \mathcal{C}_k$ (the fusion sum uses the same zero-extension); the vision adapter testbed of Appendix H uses this full form (MetaGateMLP), while the language pilot instantiates G as the posterior of the owner-supervised MLP router over concatenated logits—a simplification adopted for scale. The two instantiations have not been compared head-to-head on one testbed; within the language pilot, however, the soft and hard variants of the simplified gate track each other within about one point at every measured K, and in the vision adapter testbed the full gate and a single-scalar temperature variant land at the same final accuracy (paired temperature − gate 0.44 ± 0.38 points over 3 seeds)—these controls show the observed

Table A3: Pure competition-size control (E3): the 75-partition experts and the test inputs of the first 15 tasks are held fixed; only the candidate set grows—no retraining, no repartition (plain-substrate raw-confidence routing accuracy %; final accuracy tracks routing within 0.1 since the two-intent experts are near-ceiling on owned inputs). For the random-distractor rows, five distractor draws are averaged within each training seed, after which mean±standard deviation is computed over the three training seeds. Per-seed paired $15 \to 75$ changes: $-1.0/-3.9/-2.9$ (mean -2.6 ± 1.5). All differences are computed before rounding.

Candidate experts	15	30	50	75
Prefix (arrival order)	91.3±3.3	90.9±3.4	89.8±3.3	88.7±3.3
Random distractors	—	90.7±3.1	89.9±3.2	—

budget sensitivity is not unique to the one-hidden-layer MLP, although we do not claim independence from router parameterization.

- *Prototype routing* (nonparametric: centroids are estimated from data but nothing is fitted across tasks): route to $\arg\max_k \cos\big(e(x), \mu_k\big)$ over frozen-embedding task centroids μ_k.
- *Readouts on RAS-trained adapters* (unfitted): *RAS training* adds the shared-reference term $\lambda_R \, \mathbb{E}_{x_r \sim R} \, \mathrm{CE}(h_k(x_r), \texttt{slack})$ to every adapter's objective (Section 5.2; the Huber anchor and hard-negative emphasis are ablations, Table A6). The resulting adapters are scored either by the raw-competition quantity $s_k(x)$—*anchored confidence*, the classification default—or by the class–slack logit margin $e_k(x) = \max_{c \in \mathcal{C}_k} z_{k,c}(x) - z_{k,\texttt{slack}}(x)$—*anchored margin*, a scoring ablation in classification and the score used in generation. Nothing is fitted across adapters and no adapter is refitted as the library grows.
- *Generation margin competition* (unfitted): let $P_k(x)$ prepend expert k's task definition to input x, and let $z_{k,v}(P_k(x))$ be the next-token logit for token v. With v_{slack} the first token of the designated slack target,

$$e_k^{\text{gen}}(x) = \max_{v \in \mathcal{V} \setminus \{v_{\text{slack}}\}} z_{k,v}(P_k(x)) - z_{k,v_{\text{slack}}}(P_k(x)), \qquad k^* = \arg\max_k e_k^{\text{gen}}(x).$$

  Only the selected expert decodes the output (Section 6.5).
- *Oracle*: ground-truth owner routing; the ceiling that isolates arbitration loss from adapter quality.

E Additional Figures and Tables

Reading Table 2. The raw competition does not use the router and is constant by construction. With adapters fixed, end-to-end accuracy falls from 86.8 to 53.7 at B_r=100, and to 37.3 at B_r=50. At budget 100, jointly reducing the adapter-training buffer yields 35.4, compared with 53.7 when adapters are held fixed—a *conditional joint-starvation* effect, not an additive decomposition. At B_r=300 the router (76.1 ± 1.5) already trails the raw competition (85.2 ± 1.0) by 9.0 ± 1.0 points. The logistic-regression control recovers ∼7 points at low budget but still collapses and stays far below raw (60.6 vs. 85.2 at B_r=100): the collapse is not specific to the one-hidden-layer MLP. Per-seed MLP values: B_r=1000 $\{86.9, 86.9, 86.6\}$, 300 $\{74.6, 77.6, 76.2\}$, 100 $\{52.2, 53.6, 55.2\}$, 50 $\{39.1, 37.6, 35.1\}$.

E.1 The full reference-anchored grid with hard-negative mining

Table A5 reads the whole grid for the hard-negative-mining ablation (reference plus mining), scored by anchored confidence; the default protocol adds no hard negatives. It lifts the confidence competition in all seven cells and in every seed, most where the library is large; on matched arms (mining off on both sides, same seeds and partitions) the reference alone is worth $+2.33 \pm 1.13$ in the starved cell and $+6.28 \pm 3.88$ at K=75, while mining adds $+0.19 \pm 0.03$ and -0.87 ± 1.63 (Table A6). Anchored confidence holds 87.2 ± 0.6 to 83.7 ± 1.9 across a 20× budget and 5× library range, against 85.3 ± 1.3 to 78.2 ± 4.0 without the reference and a 52.8-point span for the plain fitted rules (Table 1). Replacing confidence by the class–slack margin on those identical adapters does not pay: it is the flatter readout (cell range 1.7 vs. 3.6) but trails confidence in all four budget cells

Table A4: Accuracy levels behind Table 4 at 7B–32B (final accuracy %, mean±sample SD over 3 paired seeds; hardware mixed across one H200 and RTX 4090s, partition pairing exact). Plain-margin scoring is poor at all three scales (paired Δ_K: $-22.2_{\pm 19.2}/-25.4_{\pm 8.9}/-35.5_{\pm 7.3}$, high-variance at 7B) while the anchored margin is flat ($-0.3/-0.9/-0.5$).

	7B		14B		32B	
Substrate, score	K=15	75	K=15	75	K=15	75
Plain, confidence	89.5±0.7	70.1±1.6	91.2±0.2	83.6±0.7	91.2±0.9	81.2±1.5
Anchored, confidence	91.7±0.5	90.2±0.4	93.3±0.2	91.5±0.5	93.0±0.4	91.4±0.9
Plain, margin	31.9±18.0	9.7±3.4	43.7±5.1	18.4±8.4	51.7±7.8	16.2±4.7
Anchored, margin	90.3±0.3	89.9±0.5	92.2±0.1	91.3±0.5	91.8±0.7	91.3±0.9

Table A5: Reference training and scoring readout on the CLINC150 grid (final accuracy %, mean±sample SD over 3 seeds, identical class partition per seed; bold marks the confidence readout, the classification default). Reference training substantially improves the confidence competition at large K, whereas replacing confidence by the class–slack margin provides no reliable additional gain and is harmful in the small-library budget sweep—the margin is nominally ahead only at K=75, by $+1.30 \pm 1.62$ per seed. Plain row: the no-reference substrate of Table 1, extended with a B=50 cell. The reference effect isolated on matched, mining-off arms is in Table A6.

	Budget axis (K=15)				Library axis (B=1000)		
	B=50	100	300	1000	K=30	50	75
Plain confidence (no reference)	83.6±1.2	83.9±1.1	85.3±1.3	85.2±1.0	83.0±1.2	80.9±2.5	78.2±4.0
Anchored confidence (+ mining)	**85.1**±0.2	**86.4**±0.4	**87.1**±1.3	**87.2**±0.6	**85.4**±1.2	**84.1**±1.6	**83.7**±1.9
Anchored margin (+ mining; ablation)	83.3±0.5	83.5±1.0	84.2±0.4	83.3±0.4	84.9±0.7	83.7±1.7	85.0±1.2

(-3.90 ± 0.67 at B=1,000, -2.85 ± 0.93 starved) and turns positive only at K=75 ($+1.30 \pm 1.62$, one seed driving it). On the resource-symmetric control, the reference's 250 samples given as replay instead (B=350), plain confidence reaches 85.2 ± 1.8 (a within-seed gain of 1.37 ± 0.75) and the anchored starved cell keeps a paired advantage of 1.16 ± 1.86 points.

Coordinate check for Figure 2(b). The τ decline is not an artifact of the margin coordinate: in the confidence coordinate the K=75 tail falls identically ($0.95 \to 0.81 \to 0.49$ across 125M→1.5B, then plateaus); at K=15 softmax saturation pins that coordinate near 1.0 at every scale, which is why τ is reported in margin space. Both components also move favorably: $q_{0.99}(S^-)$ falls ($2.7 \to -0.1$, 125M→1.5B) while the owner–wrong gap $\mu_+ - \mu_-$ widens ($15.2 \to 19.0$).

E.2 Reference geometry and the budget-ratio control

The reference-swap arms of Section 6.3 also separate the two readouts. The margin ablation depends on the reference far more strongly than the confidence competition does: without a reference it collapses (29.3 ± 3.4 on banking77, 20.3 ± 7.9 on FewRel—45 and 27 points below the confidence read from the same adapters), a far reference recovers part of it (60.9 ± 2.9, 30.7 ± 7.1), and only an adjacent one brings it level with confidence (79.6 ± 1.7, 59.7 ± 1.4). Its swap cost exceeds the confidence's in every seed on both benchmarks, so a shared coordinate is only as comparable as the reference that defines it.

A budget-ratio consistency check varies K and the per-route sample count together: at an identical 13 buffer samples per route, the fitted router scores $88.2 \pm 0.7\%$ when its *total* meta-set is 975 (K=75, B=1,000) but only $60.0 \pm 1.5\%$ at 195 (K=15, B=200; 3 seeds). This is consistent with total meta-data rather than the per-route ratio governing; the two cells also differ in K and class granularity, and the fixed-adapter, fixed-K control of Table 2 isolates the budget effect directly.

F Limitations

The map's cells are delimited by measurements: the fitted gate wins by up to $+5.30 \pm 0.35$ points at K=2 in the vision adapter testbed (Appendix H), the CLINC starvation threshold sits between 300 and 1,000 total samples at K=15, and the difficulty axis is anchored by banking77 (hard) and

Table A6: Training × readout at the two CLINC150 decision cells (final accuracy %, mean±sample SD over 3 seeds, identical class partition per seed; the owner-in-library oracle is 99.1 ± 0.1 starved and 99.8 ± 0.0 at K=75 and moves by less than 0.4 of a point in every seed on the reference, mining and hybrid separate-head rows, while the margin-anchor rows give up 0.4 and 0.8 points of that ceiling, adversely in every seed; bold marks the deployed readout, RAS training read out by confidence). Top block, the 2×2 core: reference-as-slack training is the decisive component; margin scoring is not required for classification and is the weaker readout in the starved cell. RAS training read out by confidence is *anchored confidence*; read out by the margin, *anchored margin*. Bottom block: recipe variants, the separate competence head—the only setting in which an auxiliary competence head is trained—and the fitted soft gate on the mining-recipe adapters, for context.

Training	Readout	K=15, B=100	K=75, B=1000
plain (no reference)	confidence s_k	83.9±1.1	78.2±4.0
plain (no reference)	margin e_k	43.9±8.2	69.5±5.3
RAS training (reference as slack)	**confidence** s_k	**86.2**±0.5	**84.5**±0.3
RAS training (reference as slack)	margin e_k	83.5±0.3	84.3±1.1
+ hard-negative mining	confidence s_k	86.4±0.4	83.7±1.9
+ hard-negative mining	margin e_k	83.5±1.0	85.0±1.2
+ margin anchor (τ_R=−2)	confidence s_k	84.1±0.8	80.8±0.7
+ margin anchor (τ_R=−2)	margin e_k	80.7±0.9	80.5±0.9
separate competence head (hybrid)	head score g_k	67.7±0.4	82.7±0.6
separate competence head (pure)	head score g_k	63.2±3.3	83.1±1.3
soft gate (fitted, mining adapters)	—	47.1±1.4	88.8±0.7

Table A7: CLINC150 pilot (frozen RoBERTa-base, LoRA r8 per task, buffer 1,000, mean±sample SD over 3 seeds): final accuracy (%) of four arbitration rules as the library grows. The oracle is flat; prototype routing collapses; the fitted MLP router never meaningfully beats the unfitted abstention competition (within-seed paired lead $+0.9$ to $+1.4$ points at $K \leq 8$, each inside one sample SD except K=2) and ends below it at K=10 (-1.2 ± 1.1).

K	2	4	6	8	10
Oracle (ceiling)	99.0±0.3	98.7±0.7	98.7±0.4	98.8±0.4	98.7±0.4
Raw competition (unfitted)	96.7±0.8	93.8±2.5	90.7±2.5	88.2±1.0	**87.0**±1.0
MLP router (fitted)	**97.6**±0.6	**94.9**±0.9	**92.1**±1.6	**88.8**±0.7	85.8±0.3
Prototype routing	69.9±3.6	46.8±2.2	34.9±2.3	27.2±1.3	23.8±0.9

CLINC/SNI (easy) rather than measured continuously; the budget-ratio control points to total metadata rather than the per-route ratio (Appendix E.2). Slack training requires foreign negatives at adapter-training time (a replay buffer or peer data); fully isolated training is out of scope, richer repair combinations remain open, and task-boundary-free *training* streams are future work (inference is already task-free throughout)—the abstention signal needs no boundaries, but per-task adapter allocation does; the competition's $O(K)$ inference cost is measured in Appendix J.

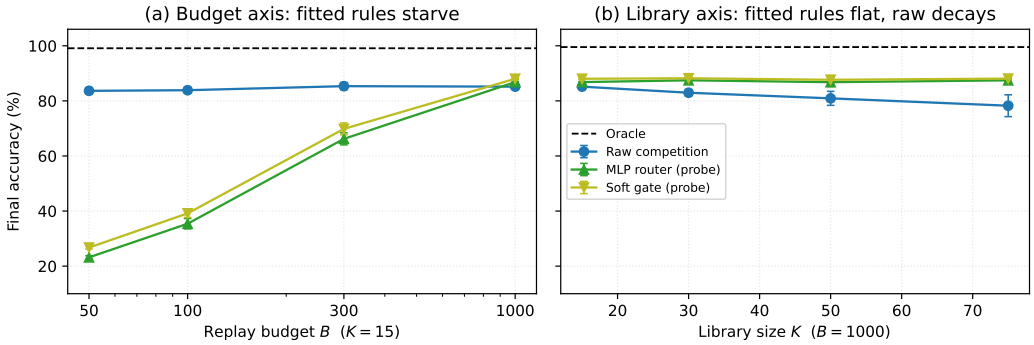


Figure A1: The two-resource pattern on the plain substrate (3-seed means; all rules on identical adapters, Table 1 plus B=50). (a) Fitted rules collapse as the replay budget shrinks while the raw competition stays far more stable. (b) A larger library (finer repartition of the fixed intent set) leaves fitted rules intact (total meta-data constant, routing easy) but erodes the raw competition through cross-expert comparability noise. The oracle never moves on either axis.

Table A8: Original published systems (as released) vs. our rules on CLINC150 (final accuracy %, mean±sample SD over 3 seeds; the mechanism-probe row shows 3-seed means as a range). O-LoRA is a sequential single model over the 15-task stream, trained in each seed on the same class-to-task partition as our rules. The task-aware row reveals task identity through per-task label options and is oracle-grade by construction. Our two rows are labelled by substrate—the raw competition is the plain substrate (λ_R=0), the anchored row is the default protocol (RAS training without hard negatives); bold marks the deployed readout, RAS training read out by confidence, not the best entry in the column (the fitted probe is higher at K=75).

	K=15, B=100	K=75, B=1000
AdapterFusion (original code)	11.5±0.3	50.2±2.4
O-LoRA (original code, class-IL)	62.0±4.7	–
O-LoRA (original code, task identity in prompt)	*98.1*±0.5	–
Fitted mechanism probe (router/gate; 3-seed means)	35.4–39.2	87.5–88.1
Raw competition (plain substrate)	83.9±1.1	78.2±4.0
Anchored confidence (RAS-trained)	**86.2**±0.5	**84.5**±0.3

Table A9: The backbone axis (CLINC150, RAS training with λ_R=0.25 and no hard negatives at every scale; anchored confidence readout; final accuracy %, mean±std over 3 seeds; Δ is the within-seed paired difference, formed before rounding). Left: the starved cell reproduces the fitted collapse at every scale. Right: the library axis at fixed B=1,000—the anchored confidence competition's K-cost Δ is shallow at every backbone.

	K=15, B=100 (starved)			B=1,000 fixed				
Backbone (frozen)	Router	Gate	Conf.	Router$_{K15}$	Gate$_{K15}$	Conf.$_{K15}$	Conf.$_{K75}$	Δ
RoBERTa-base (125M)	39.8±1.4	45.1±0.8	86.2±0.5	87.3±0.5	88.2±0.2	87.0±0.2	84.5±0.3	−2.5±0.4
Qwen2.5-0.5B	21.8±0.8	34.7±1.1	80.4±0.2	67.8±0.5	77.9±1.2	83.8±1.3	81.6±0.4	−2.2±1.0
Qwen2.5-1.5B	27.0±1.2	41.3±2.4	88.0±0.9	79.5±1.1	86.7±1.3	89.4±1.0	89.3±0.5	−0.1±0.6
Qwen2.5-3B	30.6±1.1	45.2±1.5	90.3±0.8	84.4±1.4	90.1±1.5	91.5±0.8	90.6±0.2	−0.8±1.0
Qwen2.5-7B	29.3±1.3	43.6±0.9	90.2±0.5	83.0±1.0	89.7±0.8	91.7±0.5	90.2±0.4	−1.5±0.7
Qwen2.5-14B	33.5±1.6	47.4±0.8	92.4±0.7	88.0±0.4	92.4±0.5	93.3±0.2	91.5±0.5	−1.9±0.3
Qwen2.5-32B	34.9±0.6	47.6±1.0	92.8±0.3	87.9±1.0	92.2±0.5	93.0±0.4	91.4±0.9	−1.6±0.9

Table A10: Library growth $K{=}50 \rightarrow 75$ (CLINC150, $B{=}1000$, RAS-trained adapters without hard-negative mining; mean±sample SD over 3 seeds, changes paired within seed; bold marks the deployed readout, RAS training read out by confidence): old-task accuracy before and after growth, and per-growth arbitration cost. "Old tasks" are the first 50 tasks of the 75-task partition (2 intents each)—a finer partition than the flat $K{=}50$ grid cell (3 intents each), so levels are not comparable across tables.

	old-task acc. @$K{=}50$	@$K{=}75$	refit needed
Anchored confidence	**88.7**±0.4	**87.5**±0.5 (−1.20±0.09)	none
Anchored margin (ablation)	88.1±1.3	87.1±1.3 (−1.00±0.40)	none
MLP router (refitted)	91.7±0.7	86.6±0.4 (−5.03±0.45)	every growth step

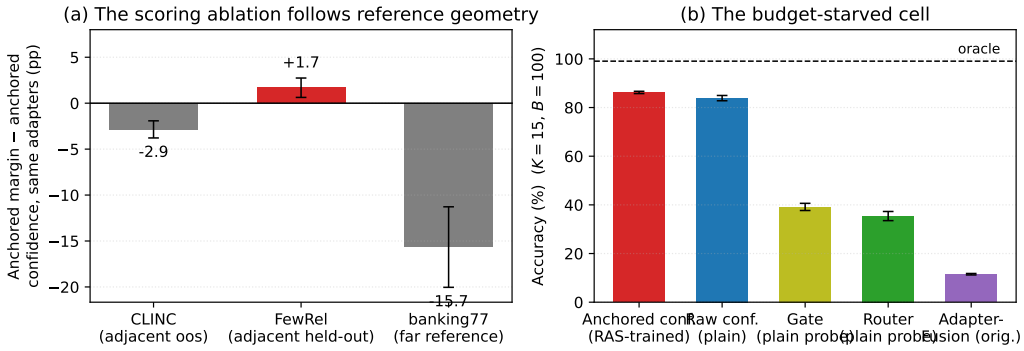


Figure A2: (a) The scoring ablation by benchmark (anchored margin minus anchored confidence on identical RAS-trained adapters with hard-negative mining, the recipe of the banking77 and FewRel main streams; starved cells, 3 seeds): the margin readout costs $2.9{\pm}0.9$ points on CLINC and $15.7{\pm}4.4$ on banking77, whose reference is far, and adds 1.7 ± 1.1 on FewRel—its sign follows reference geometry, which the within-benchmark swaps of Section 6.3 measure directly. (b) The budget-starved cell ($K{=}15$, $B{=}100$; anchored confidence under the default protocol): unfitted competitions lead, the fitted probes starve, and the original published fitted system starves hardest.

Table A11: FewRel relation-extraction stream ($K{=}16$, RAS-trained adapters with hard-negative mining, final accuracy %, mean±sample SD over 3 seeds; bold marks the best rule in each row, the oracle aside): with a domain-adjacent reference (the stream's held-out relations), the two readouts on these identical adapters are within a point at the full budget, where the fitted rules lead, and the anchored margin is the best rule under starvation—the one classification cell in which the margin readout beats confidence. Every row here compares readouts and fitted rules on one substrate; the matched no-reference arm that prices the training protocol is reported in the text (Section 6.3) and was run on the same machine.

	Anchored conf.	Anchored margin	Router	Gate	Oracle
$B{=}1000$	64.8 ±3.4	63.7 ±2.7	70.7 ±0.8	**71.8** ±1.0	96.6 ±0.6
$B{=}100$	57.9 ±1.4	**59.6** ±0.7	44.2 ±1.7	47.5 ±1.2	95.5 ±0.6

Table A12: LLM-scale decision cells (Qwen2.5-1.5B, frozen, per-task LoRA; RAS-trained adapters ($\lambda_R{=}0.25$, no hard negatives) scored two ways; final accuracy %, mean±std over 3 seeds; oracle column: means).

	Anchored conf.	Anchored margin	Router	Gate	Oracle
$K{=}15$, $B{=}100$	**88.0**±0.9	86.4±0.8	27.0±1.2	41.3±2.4	99.1
$K{=}75$, $B{=}1000$	**89.3**±0.5	89.1±0.4	83.0±0.8	88.8±0.2	99.8

Table A13: Generation cells (Super-NaturalInstructions, $K=20$, frozen Qwen2.5-1.5B, winner-only decoding; routing accuracy % / system Rouge-L F1; mean±sample SD over 3 seeds, each seed a fresh task draw; no-reference $B=1{,}000$: single run, reported without a SD). The no-reference row is the same first-token margin without RAS training, broken by the first-expert asymmetry.

	$B=100$ (starved)	$B=1{,}000$
First-token margin, no reference	7.5±3.8 / 9.5±2.9	7.8 / 8.7
First-token confidence (no slack)	67.2±17.5 / 53.7±11.9	75.1±9.7 / 57.6±3.9
MLP router (fitted)	**88.7**±3.7 / 60.2±6.7	**96.9**±0.9 / **63.5**±4.0
RAS training + first-token margin	85.3±7.4 / 60.2±6.5	85.8±5.8 / 59.7±5.0
Oracle	100±0.0 / 66.9±4.8	100±0.0 / 65.0±4.5

G PUBLISHED SYSTEMS ON THE IDENTICAL STREAM

Our fitted rules are minimal *mechanism probes*; we also ran two original implementations on the identical CLINC 15-task stream: **AdapterFusion as released** (two-stage `adapters`-library procedure, fusion trained on the replay buffer—the same cross-task data every fitted rule receives; Table A8) and **O-LoRA as released** (Wang et al., 2023) (official code, T5-large, orthogonal-subspace LoRA, no replay), representing the single-model LoRA continual-learning line.

AdapterFusion lands far below even our probes in both cells (11.5/50.2%): its fusion attention and global head fit far more parameters on the same limited data. Replacing the probes by the published systems we could obtain widens the unfitted lead (anchored confidence over AdapterFusion: $+74.7$ and $+34.3$): every published fitted system we ran does worse on the matched task-free stream, and its collapse on original code is evidence about the protocol—fitted components starve on replay budgets. O-LoRA makes the complementary point: under its native protocol, where each test prompt carries its own task's ten label options (task identity revealed), it reaches 98.1%—oracle-grade; withhold task identity by offering all 150 seen labels and the same model drops to 62.0%, 36 points lost entirely to arbitration, on a backbone six times larger than the RoBERTa library whose raw competition scores 83.9%. The published LoRA line is strong precisely because its protocol removes the problem we study; once the problem is posed, a library with the raw competition beats the sequential single model by over twenty points.

H VISION STRESS TESTS: TWO SEPARATE TESTBEDS

Our vision evidence comes from two separately implemented systems that share only the dataset (CIFAR-100 (Krizhevsky, 2009)). **Neither establishes vision transfer of RAS training.** Testbed (a) shows that a disjoint reference leaves anchored confidence unchanged and does not support the anchored margin on frozen vision adapters, and testbed (b) shows a performance loss when RAS training is applied to a representation that is still being trained. All entries from the two testbeds are means $\pm$ sample standard deviation over seeds 1–3; paired differences are formed within seed, never across the two testbeds and never across host groups.

H.1 TESTBED (A): FROZEN-BACKBONE ADAPTER LIBRARY

What it is. An ImageNet-pretrained ResNet-18 (He et al., 2016) (torchvision `IMAGENET1K_V1`, native 7×7/stride-2 stem) frozen from task 0—weights and batch-norm statistics verified bit-identical to the pristine checkpoint at every task boundary—carrying one LoRA expert per task (r=4, α=16, layer3–4 convolutions) with a 101-way head (100 classes + slack) trained by cross-entropy over all 101 outputs; own-class probabilities are read from the restricted 101-way softmax. Split CIFAR-100 10×10 in fixed class order, reservoir buffer, per-step 1:1 slack replay, and no expert trained after its own task (no retraining, no guard). Every arbitration rule is scored on the *same* experts at every K and seed.

What it shows. The fitted gate's routing accuracy decays by -4.68 ± 0.02 points per added expert (OLS over K=2–10) while the oracle ceiling stays within 82.56–83.59 (Table A14); at K=10 oracle routing exceeds the unfitted competition by 40.84 ± 0.27 points, so the experts are capable and the loss is arbitration. The unfitted competition overtakes the best fitted alternative from K^*=8 (per-seed persistent crossover $[7, 8, 8]$): the gate leads only at small K (maximum paired advantage $+5.30 \pm 0.35$ at K=2, crossing zero between K=7 and K=8), while prototype routing and *class-wise trust*—a train-free rule that estimates on the buffer a Beta$(1, 1)$-smoothed precision for every (expert, class) claim, certifies the experts whose precision clears the prior by a fixed margin (each class keeping at least its best expert), and then takes the maximum own-class probability over certified experts only—are overtaken already at K=2. Starving the shared buffer reproduces the language budget axis here: the paired gate-minus-competition difference falls from -2.53 ± 0.61 at B=1,000 to -8.79 ± 0.51 at B=500 and -17.49 ± 1.38 at B=200 (competition levels 42.27 ± 0.43, 40.53 ± 0.54, 36.87 ± 0.32). Final accuracies at K=10, B=1,000 on identical experts: unfitted competition 42.27 ± 0.43, temperature-calibrated max 40.19 ± 0.44 (per-expert buffer-fitted temperature, Appendix C.1), fitted gate 39.75 ± 0.53, class-wise trust 18.89 ± 2.49 (the owner-tie variant reads 20.66 ± 3.14; the fraction of classes whose owning expert is certified falls from 0.683 at K=2 to 0.243 at K=10), prototype routing 18.00 ± 0.14 under the task-centroid rule of Appendix D (the

Table A14: Vision testbed (a), the frozen-backbone adapter library (ImageNet-pretrained ResNet-18, per-task LoRA experts, Split CIFAR-100 10×10, B=1,000), vs. library size K: the fitted gate's *routing accuracy* decays by -4.68 ± 0.02 points per added expert (OLS), while the *final accuracy under oracle routing* is flat—the ceiling never moves. Entries are means over 3 seeds; the per-K sample standard deviation is at most 1.47 in the gate row and at most 0.90 in the oracle row. Every rule is scored on the same experts within each seed.

K	2	3	4	5	6	7	8	9	10
Fitted gate (routing acc.)	81.50	70.81	63.10	58.31	53.97	50.97	47.50	44.67	40.53
Oracle-routed final acc. (ceiling)	82.85	83.59	82.81	82.95	82.81	82.56	82.60	82.63	83.12

class-max variant implemented in our code reads 31.06 ± 0.15), oracle 83.12 ± 0.27; across the four B=1,000 adapter arms the competition ends between 37.77 ± 0.40 (with slack-mixup) and 42.27 ± 0.43 (no reference). Two rule-definition checks are immaterial: the zero-extended meta-gate of Appendix D and the gate as deployed agree (39.75 ± 0.53 vs. 39.88 ± 0.20), and the gate and the temperature-calibrated max are statistically indistinguishable (paired temperature − gate 0.44 ± 0.38, 95% t CI $[-0.49, +1.38]$).

What testbed (a) shows. Transplanting RAS training here does not reproduce the language behaviour. With the disjoint SVHN reference at λ_R=0.25, anchored confidence does not differ from the matched no-reference arm (41.98 ± 0.42 vs. 42.27 ± 0.43; paired -0.29 ± 0.47, 95% t CI $[-1.45, +0.87]$): the reference neither helps the deployed readout nor materially costs it, and reducing λ_R five-fold changes nothing either (paired $+0.09 \pm 0.32$). The anchored margin—the deployed readout in generation, a scoring ablation in classification—collapses outright (17.36 ± 0.42 final) while the same experts keep an 83.00 ± 0.08 task-incremental ceiling, a pure score-comparability failure, even though the reference does improve that readout relative to no reference ($+6.17 \pm 0.40$ paired). Slack-mixup, which interpolates *input images* rather than embeddings in our implementation, costs the competition -4.50 ± 0.37 while leaving the oracle unchanged ($+0.12 \pm 0.28$) and helping the fitted rules (gate $+1.66 \pm 0.75$, temperature-calibrated max $+1.84 \pm 0.46$). No arm varies reference proximity—SVHN, disjoint from every CIFAR task, is the only vision reference we run—so far-reference geometry remains an untested hypothesis here rather than an explanation, and an in-domain-adjacent vision reference is the concrete open direction.

H.2 Testbed (b): trainable full-network system

What it is. A separately implemented system with no adapters and no frozen backbone: one from-scratch ResNet-18 per task (101 outputs), arbitrated by the same restricted-head max-confidence competition, on Split CIFAR-100 at K=10 and K=20. Its lifecycle differs from testbed (a) in one respect that matters: a one-time *warm-start guard* re-optimizes the *first network* at the end of task 2, from its own weights at a hard-coded 0.1× learning rate, on buffered task-1 samples with true labels plus a class-balanced 1,000-sample subset of task-2 data labelled slack. No other network is ever revisited, and no LoRA adapter exists in this system. The single-network baselines (iCaRL (Rebuffi et al., 2017), DER++ (Buzzega et al., 2020), ER-ACE (Caccia et al., 2022)) run in Mammoth (Buzzega et al., 2020) under two schedules: the *Mammoth default schedule* (`seq-cifar100` dataset defaults plus one hand-set learning rate and momentum applied uniformly to all three methods) and the *matched schedule* (our data pipeline and training schedule, ColorJitter included; iCaRL keeps its own weight decay). The stack holds K networks and therefore costs 10–20× the parameters and the inference compute of a single-network baseline.

What it shows. Table A15. The stack reaches 52.60 ± 1.16 at K=10 and 43.88 ± 0.49 at K=20; with its lifecycle controls, the K=10 cells of this system span 50.14–52.60. Against the matched schedule—the only seed-paired comparison available, since the default-schedule baselines were produced on a second machine and are reported unpaired—it leads the best baseline cell by 4.61 points (seed-paired 4.61 ± 1.00, 95% t CI $[+2.12, +7.10]$). Over the full six-cell baseline grid at K=20 the best cell is a default-schedule one and the stack leads it by 6.02 points, unpaired. The stream-length ordering does not survive three seeds: from K=10 to K=20 the stack loses 8.72 ± 1.64 points, whereas under the matched schedule DER++ *gains* 6.01 ± 1.55 and ER-ACE is flat (-0.73 ± 2.36), and only matched-schedule iCaRL degrades faster (10.46 ± 0.19); under the Mammoth default schedule every baseline also loses less than the stack does (iCaRL 29.36 → 26.37,

Table A15: Vision testbed (b), the trainable full-network system (one from-scratch ResNet-18 per task, no adapters; Split CIFAR-100, final Class-IL accuracy %, mean±sample SD over 3 seeds). Top: the abstention stack against single-network baselines under two schedules; only the matched-schedule cells share a machine with the stack and admit seed-paired differences, while the Mammoth-default cells are unpaired method-vs-method comparisons. Bottom: lifecycle and regime ablations at K=10 (the no-guard cell is also the pure control for the regime ablations).

	K=10	K=20
Abstention stack (K networks)	**52.60**±1.16	**43.88**±0.49
iCaRL (matched schedule)	47.99±0.17	37.53±0.12
DER++ (matched schedule)	24.83±0.32	30.84±1.85
ER-ACE (matched schedule)	31.20±1.47	31.94±0.95
iCaRL (Mammoth default schedule)	29.36±0.20	26.37±0.10
DER++ (Mammoth default schedule)	44.98±0.85	37.62±0.90
ER-ACE (Mammoth default schedule)	41.61±1.26	37.86±1.05
no guard (= pure control)	50.82±0.37	—
no warmup	50.14±0.31	—
+ slack-mixup (input space)	47.42±0.65	—
+ SVHN reference (λ_R=0.25)	39.87±1.80	—

DER++ $44.98 \rightarrow 37.62$, ER-ACE $41.61 \rightarrow 37.86$, read off the cell means, since those cells share no machine with the stack and admit no paired difference). The ordering therefore fails under both schedules.

What testbed (b) shows. Both training-side ablations run against the method here. RAS training with the SVHN reference costs 30.67 ± 14.18 points of first-task accuracy against the matched no-reference control and ends 10.95 points lower overall (39.87 ± 1.80 vs. 50.82 ± 0.37); slack-mixup costs 3.40 ± 0.45 (47.42 ± 0.65). This testbed shows a performance loss under reference training but does not isolate its cause. Two lifecycle disclosures: the guard buys 1.78 ± 0.81 points end-to-end (95% t CI $[-0.23, +3.79]$, unresolved at three seeds) while costing the first network 24.47 ± 1.71 points of its own final task-1 accuracy; and the one-epoch warmup is worth 2.46 ± 1.42 points, though at K=20 its ramp is hard-coded to 40 steps against 20 steps per epoch and so completes only half way through the first epoch.

I Hyperparameters, Protocol, and Seed Semantics

Language classification (CLINC150, banking77, FewRel; frozen RoBERTa-base). LoRA r=8, α=16 on attention query/value, dropout 0; linear head on the [CLS] state with one slack output; AdamW, lr 2×10^{-4}, batch 32, 10 epochs per task with a one-epoch linear warmup ($0.01 \rightarrow 1$); foreign replay samples relabeled slack in every batch; task-balanced replay buffer with capacity B. For each RoBERTa expert, a uniformly sampled holdout of $\max(8, \lfloor 0.1N \rfloor)$ examples from the combined native and foreign pool of size N is excluded from adapter training and used for the per-expert mean-shift, ECDF and temperature controls. On reference-carrying substrates, each batch adds $\min(32, |R|)$ reference samples trained as slack with weight λ_R=0.25 ($|R|$=250 CLINC out-of-scope utterances, used by the CLINC grid, by the banking77 main streams and by the far arm of the banking77 swap; held-out relations for FewRel, and 7 held-out banking77 intents for the adjacent arm of the banking77 swap); Hard-negative mining is used in the full-grid recipe (Table A5) and the banking77 and FewRel main streams, but not in the plain grid, the reference-only factorial comparison (Table 3), or the reference-swap arms. From epoch 2, when at least eight foreign training examples are available, the top half by current class–slack margin are duplicated once in that epoch's training pool. The margin-anchor ablation uses Huber loss with λ_A=0.5, $\tau_R = -2$ and transition parameter δ=1. Fitted probes: scikit-learn `MLPClassifier` (one hidden layer of 128, up to 1,500 iterations) on concatenated per-adapter logits with owner labels; the soft gate reads the same posterior as a fusion weight; prototype routing uses frozen-[CLS] task centroids.

Backbone sweep (frozen Qwen2.5, 0.5B–32B, bf16). Identical LoRA and loss configuration on `q_proj`/`v_proj`; lr 2×10^{-4}, batch 16, 10 epochs per task; λ_R=0.25, no hard negatives; eager attention on the H200 runs.

Generation (Super-NaturalInstructions, frozen Qwen2.5-1.5B). K=20 tasks, 300 train / 50 test examples per task; lr 2×10^{-4}, batch 8, 3 epochs per task; reference: 100 inputs from five held-out SNI tasks trained to the designated out-of-scope literal in every expert. The slack target is a leading space followed by `IRRELEVANT`; v_{slack} is the first token returned by the model tokenizer with `add_special_tokens=False`. Prompts have the form "task definition, newline, `Input:` input, newline, `Output:`". Training loss is applied to all target tokens and EOS; prompt and padding positions are masked. The fitted generation router uses frozen-base last-token representations of the raw inputs.

Fixed-adapter controls (Table 2). Both selectors consume the identical cached concatenated-logit features and the same nested router-label subsets (same subsampling RNG and seeds; stratified by task). The L2 logistic-regression control is multinomial with the lbfgs solver, C=1.0 fixed *a priori* (no validation selection), up to 2,000 iterations, uniform class weights; feature standardization is fitted only on each router-training subset; no labels outside B_r are used. The competition-size control (Table A3) fixes the 75-partition experts and the first-15-task test inputs and grows only the candidate set; its random arm redraws distractors, not training seeds.

Competence head. The scalar competence head g_k exists only in the CR-SLoRA ablation (Appendix C.3); every other adapter is trained without it.

Compute grouping. The CLINC150 grid and all its ablations ran as one campaign on a single H200; the banking77 and FewRel streams—main arms, matched no-reference arms and reference swaps alike—ran on two local workstations, one benchmark per machine (an RTX 4090 node for banking77, an RTX 5090 node for FewRel). Every seed-paired difference we report is formed inside one campaign, never across machines.

Seed semantics. A seed on CLINC/banking77/FewRel redraws the class-to-task partition and all training randomness (batching, per-task buffer resampling); on SNI it additionally redraws the task set; for O-LoRA it redraws the same class-to-task partition as for our rules (the release fixes its own training seed); vision seeds vary training randomness on a fixed class split in both vision testbeds; in the adapter testbed the reference draw is fixed by its own generator seed and does not vary with the run seed. Remaining details (schedulers, exact data splits, reference-selection rules) are in the released scripts.

J Regime-Matched Deployment Guidance

The map translates directly into deployment guidance (Table A16).

Inference cost, measured. The competition evaluates all K adapters. On an RTX 5090 (batch 32, short utterances) this costs 0.33 ms per input per adapter on RoBERTa-base (5.0/24.5 ms per input at K=15/75) and 0.90 ms on frozen Qwen2.5-1.5B in bf16 (13.5/67.4 ms)—linear in K, as expected. Extrapolating by parameter count, a 32B backbone at K=75 costs on the order of a second per input when adapters are evaluated sequentially, so large-K LLM deployments require batched multi-adapter serving (Sheng et al., 2024) or an independently validated top-m pre-filter (Appendix C). Scenarios in which routing is surface-easy and meta-data accumulates—per-language or per-domain adapters on a frozen multilingual encoder (the MAD-X pattern; e.g., multilingual or multi-domain NER), per-tenant personalization in SaaS services, and bounded skill hubs on LLMs—sit squarely in the router's cell, and our large-K results show the router remains sound there even at K=75 provided its *total* meta-data is adequate. Conversely, unbounded class-incremental streams that exhaust a fixed replay budget, domains where the adapters' inputs are mutually confusable (our frozen-backbone vision adapter testbed), and operational settings that require *refit-free growth*—adding an adapter without retained past data or router retraining—belong to the abstention competition, which substantially outperforms the fitted probes in the measured budget-starved cell (83.9 ± 1.1 plain, 86.2 ± 0.5 with RAS training, vs. router 35.4 ± 1.9 and gate 39.2 ± 1.5; 3 seeds).

Table A16: Regime-matched arbitration: deployment guidance derived from the two-resource map. RAS-training entries apply to its measured scope—frozen language backbones—and classification always reads the resulting adapters with anchored confidence; in our within-benchmark swaps an adjacent reference is worth 4.6 points on banking77 and 11.5 points on FewRel of that readout over a far one (Section 6.3).

Scenario	Cell	Recommended arbitration
Multilingual/multi-domain adapters (NER, classification)	easy routing, data-rich	fitted router / gate
Per-tenant personalization (SaaS)	easy routing, data accumulates	fitted router / gate
Bounded LLM skill hub, composite queries	easy routing, data-rich	soft gate (blending; composite queries untested here)
Unbounded class-incremental stream, fixed buffer	budget-starved	abstention competition (confidence readout)
Mutually confusable adapters, no suitable reference	hard routing	plain confidence competition
No router refit possible (foreign-negative source still available)	any	RAS training + anchored confidence (frozen LM) / plain competition
Domain-adjacent task-free pool available (e.g., OOS logs)	starved or refit-free cells	**RAS training** + anchored confidence (frozen-language scope)