A TRAINABLE MEMORY-MOVEMENT OBJECTIVE FOR MIXTURE-OF-EXPERTS DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding a mixture-of-experts model at batch size one is bound by memory traffic, most of it expert weights moving into a small fast tier. We make the expected bytes moved per token a differentiable training term, a capacity-constrained soft residency over the router's gate mass in a memoryless (batch-marginal) and a stateful (windowed) form, and measure what it buys with the Belady floor, the offline-optimal miss count of a model's own routing traces, which no eviction policy can beat. Three findings follow. The traffic reduction grows with expert granularity, reaching an order of magnitude at 64 experts and 3.6–6.8B parameters for a perplexity cost of 7–13%, and in a trace-driven H100 replay converts to $+63$–90% offloaded decode throughput. Which loss form wins at matched quality turns on the load-balance coefficient shared by all arms; at the library default the residency objective moves $0.39\times$ the misses of a switch-only objective at the 25% tier and $1.13\times$ at 12.5%, the residency term alone carries the effect, the stateful form leads on both axes at 25%, and a decomposition of cache misses into a marginal factor and a temporal excess explains the reversal. Under the customary retrofit protocol a model pretrained under the objective reaches a floor two to three times below a finished model retrofitted with it, across three geometries; a retrofit at the pretraining learning rate reaches the born floor at lower perplexity, and at matched total compute the two protocols tie at one balance coefficient while the born model keeps a $1.2\times$ floor lead at the other on one run, so the operating point is set by the optimization protocol more than by when pressure starts.

1 INTRODUCTION

A mixture-of-experts (MoE) language model (Shazeer et al., 2017; Fedus et al., 2022) decoding one sequence at a time is limited by bytes moved, not by arithmetic. Only a few experts per layer are active for each token, but which ones changes from token to token, and when the expert weights do not fit in the fast memory tier every change is a transfer. On OLMoE-1B-7B (Muennighoff et al., 2025), whose 1024 expert weight blocks are ≈ 12.6 MB each, a least-recently-used (LRU) cache holding a quarter of the blocks moves 0.8 GB of weights per token.

Systems work learns cache and prefetch policies for a routing workload it treats as given; locality-aware MoE training reshapes the router but evaluates it against a fixed policy. This paper puts the coupled quantity, expected memory movement per token, into the training loss and asks what sets the operating point the objective can reach. Our instrument is the Belady floor (Belady, 1966), the miss count of the offline-optimal policy on a model's own recorded routing traces, which lower-bounds every online eviction policy, so a difference in floors is one no cache design can recover. We measure it for OLMoE-architecture models of 62M to 6.8B parameters pretrained under the objective ("born"), against finished models retrofitted with it (continued training under the objective), and against competing loss forms, at three learning rates, three onset times and three load-balance coefficients.

Our contributions are three.

1. **A trainable memory-movement objective** (Section 3). A differentiable expected-bytes-moved term built on a sum-constrained soft top-M residency operator, in a memoryless (batch-marginal) and a stateful (windowed) form, trained with a stop-gradient surrogate

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.

through the capacity threshold. The sum constraint makes the term, evaluated on the access marginal, an upper bound on the miss rate of the best static pin; training evaluates it on the router's gate-mass marginal, an empirical proxy. Static tier placement, chunk-sticky consistency and frozen-residency offload objectives are limits of it.

2. **Its quality–traffic trade-off, and what carries it** (Sections 5.1, 5.2 and 5.4). The realized LRU reduction grows with expert granularity ($3\times$ at 32 experts, $6.6 \times \pm 0.5$ at 64, 14–$16\times$ at 64 experts and 3.6–6.8B), holds at $4\times$ the training tokens and on out-of-domain traces, and in an H100 trace replay converts to $+63$–90% decode throughput at a 12.5% fast tier. Against a switch-only objective the ordering reverses with the load-balance coefficient that every arm shares; at the library default and matched perplexity ours moves $0.39\times$ the misses at the 25% tier and $1.13\times$ at 12.5% (three seeds per arm), the residency term alone reproduces this, the stateful form leads on both axes at 25%, and the decomposition $\mathrm{LRU} = \mathrm{Static} \times \kappa$ accounts for the reversal. The coefficient dependence is reported in full, including that the main runs used an unintended value of 1.0.
3. **Protocol, not onset, sets the born–retrofit gap** (Section 5.3). Under the customary retrofit protocol, a peak learning rate of a quarter of the pretraining peak, the born model's Belady floor is 2.0–$3.1\times$ below the budget-matched retrofit's across three geometries. A retrofit at the pretraining peak reaches the born floor at 32 and 64 experts, 6 and 8 perplexity points lower, and a born model given the same total compute ties it at one balance coefficient and leads by $1.2\times$ at the other (one seed each). Born and retrofit floors fall on one descriptive line against the learning rate integrated over the tokens under pressure (r^2=0.78).

Appendix A records which comparisons had their predictions and reading rules fixed before the runs and which are exploratory.

2 Related Work

Locality-aware MoE training. HMoE (Wang et al., 2024), Cache-Conditional Experts (Skliar et al., 2025), BlockFFN (Song et al., 2025), StickyMoE (Kayyam, 2026) and TriRoute (Balashov & Ponomarova, 2026) reshape or co-train routing against a fixed cache policy or cost model. Closest to us is MELINOE (Raje et al., 2026), which fine-tunes a router (with LoRA adapters) under a differentiable cache-simulation penalty whose soft cache state is an exponentially decayed, capacity-normalized history of top-k requests; our stateful form is the same idea with a sum-constrained residency, and our memoryless form and the born-versus-retrofit question are what this paper adds. Propositions 2 to 4 place the static, chunk-sticky and frozen-residency objectives as limits of ours; we do not claim the first locality-aware training. ReMoE (Wang et al., 2025) is a differentiable-routing precedent with a sparsity regularizer and no locality term. **Learned cache policies.** KVP (Moschella et al., 2026) and SAECache (Fang et al., 2026) learn eviction for a frozen model, and SpecMD (Hoang et al., 2026) studies expert prefetching while flagging policy interactions as uncharacterized. **Performative prediction and continuation.** That a deployed model's stable points are generally suboptimal, and that optimality is reachable by optimizing through the induced distribution, are established (Perdomo et al., 2020; Miller et al., 2021; Izzo et al., 2021); ours is the fully controllable case, the distribution map being our own router, which is why every arm is scored on traces re-recorded after training. The temperature annealing is graduated optimization (Blake & Zisserman, 1987; Hazan et al., 2016).

3 Residency-aware training

3.1 The objective

A model with parameters θ routes each token x over expert weight blocks $e \in \{1, \dots, E\}$ of a layer, with block sizes s_e. We write $p_\theta(e \mid x)$ for the router's softmax over all E experts; with top-k routing this is a differentiable surrogate for the access indicator, whose expectation the recorded traces measure, not the inclusion probability itself. A fast tier holds M blocks and a miss costs c_{miss} per byte. We train

$$\mathcal{L} \;=\; \mathcal{L}_{\mathrm{LM}}(\theta) \;+\; a\,\mathcal{L}_{\mathrm{bal}}(\theta) \;+\; \lambda(t)\,\mathbb{E}_x\Big[\textstyle\sum_e p_\theta(e \mid x)\, s_e \,\big(1 - \rho_e(u)\big)\, c_{\mathrm{miss}}\Big], \tag{1}$$

where $\mathcal{L}_{\text{bal}}$ is the standard Switch-style load-balance auxiliary with coefficient a and $\rho(u) \in [0,1]^E$ is a capacity-constrained soft residency, a sum-constrained soft top-M over a usage statistic u:

$$\rho_e(u) \;=\; \sigma\big((u_e - b_\tau(u))/\tau'\big), \qquad \textstyle\sum_{e'} \rho_{e'}(u) = M, \tag{2}$$

with scale $\tau' = \tau(t)\,\mathrm{std}(u)$ and temperature $\tau(t)$ annealed toward 0; the threshold $b_\tau(u)$ is the unique solution of the sum constraint, found by bisection, so $\sum_e \rho_e(u) = M$ at every temperature and ρ converges to the hard top-M indicator as $\tau \to 0$. Both $b_\tau(u)$ and $\mathrm{std}(u)$ are held constant under differentiation, so the update is a stop-gradient surrogate that omits the implicit term $-\partial b/\partial u_j = -q_j / \sum_k q_k$, $q_i = \rho_i(1-\rho_i)$; the forward value, on which Proposition 1 rests, is exact, and Appendix A.1 lists both passes with every stop-gradient marked. With $s_e = s$ and $c_{\text{miss}} = 1$ the bracket in Eq. 1 is the fraction of a token's routing mass that lands off the soft-resident set.

Two usage statistics. The **memoryless form** takes $u_e = \bar{p}_\theta(e)$, the current batch marginal of $p_\theta(e \mid x)$. The **stateful form** takes, per sequence, an exponential moving average of the soft accesses of the strict past, $u_{t,e} = \sum_{k \geq 1} \frac{1}{W}(1 - \frac{1}{W})^{k-1} p_\theta(e \mid x_{t-k})$ with window W and $u_1 = 0$, so token t is scored against a residency built from what preceded it; u carries gradient into the past tokens' routing, and b_τ and $\mathrm{std}(u)$ stay frozen. The second term of Eq. 1 is then the gate-mass proxy for the bytes moved per token when the fast tier holds the soft top-M blocks by u. Unless stated otherwise, arms labelled "ours" add the adjacent-token switch penalty of Section 3.3 at half weight; "residency-only" arms omit it. The loss uses a per-layer budget M_ℓ; capacities in evaluation are total block counts $M = L\,M_\ell$ and are quoted as a fraction of all blocks.

Schedules. The dose $\lambda(t)$ ramps linearly from zero to the value every table reports, reached at the last step; in born runs the ramp begins after 20% of training, in retrofits at the first step, and $\tau(t)$ anneals geometrically from 0.5 to 0.05 over the same span. Because training changes the routing distribution the residency was fitted to, every arm is scored on traces re-recorded after training, never on the traces its loss saw.

3.2 What the two terms bound

Write A for the accesses per token (layers $\times$ top-k) and N for the accesses in a trace, LRU_M and Belady_M for the miss counts of LRU and of the offline optimum at capacity M, Static_M for the miss count of a fixed pin of the M most frequently accessed blocks, $\bar{a}$ for the empirical access marginal (the frequency with which each expert enters the top-k, which the traces measure) and $\bar{p}$ for the batch mean of the gate mass, which the loss consumes; the two can differ (two top-1 tokens with gate rows $(0.99, 0.01)$ and $(0.49, 0.51)$ have $\bar{p} = (0.74, 0.26)$ and $\bar{a} = (0.5, 0.5)$).

Proposition 1 (the miss term upper-bounds the static pin)**.** *For unit sizes, any $\tau > 0$, and $\bar{a}$ the empirical access marginal of a trace, $\sum_e \bar{a}_e\,(1 - \rho_e(\bar{a})) \;\geq\; 1 - \mathrm{Top}_M(\bar{a}) \;=\; \mathrm{Static}_M/N \;\geq\; (\mathrm{Belady}_M^{\mathrm{byp}} - M)/N$, where $\mathrm{Belady}_M^{\mathrm{byp}}$ is the offline optimum when a miss may be served without insertion. The training term evaluates the left-hand side at $\bar{p}$, for which no such bound holds (the two-token example gives 0.26 against a static miss fraction of 0.50), so its use in training is an empirical connection.*

The first inequality holds because a linear functional over $\{0 \leq \rho \leq 1,\ \sum \rho = M\}$ is maximized at a hard top-M indicator, and fails for an unconstrained sigmoid; the second because a pin is a bypass policy. The demand-paging Belady we report lower-bounds every demand policy but can exceed $\mathrm{Static}_M + M$ in general (capacity one, trace a, b, a, b); on all 136 traces in this paper it satisfies the bound with the fill charged per chunk. Both marginals are invariant to reordering tokens, so the miss term targets Static_M and is blind to the temporal excess $\kappa = \mathrm{LRU}_M/\mathrm{Static}_M$; the adjacent-token switch count $\mathrm{Sw}_1 = \sum_t |S_{t+1} \setminus S_t|$ upper-bounds LRU misses beyond the first token's fill whenever $M \geq 2A$, and neither quantity bounds the other. For $M \leq A - k$, with k the per-layer top-k, LRU misses on every access, so no loss form can move LRU at a tier that cannot hold one token's footprint, the 6.25% tier of our geometries. Proofs are in Appendix B.

3.3 Existing objectives as limits of Eq. 1

As $W \to \infty$ under i.i.d. accesses the stateful objective becomes a static tier-placement penalty on non-resident routing mass (Proposition 2), the class of fixed-placement and cache-conditional methods (Skliar et al., 2025; Wang et al., 2024). With hard residency taken as the blocks used in the pre-

vious chunk, which requires relaxing the sum constraint, it becomes a chunk-sticky consistency objective (Song et al., 2025; Kayyam, 2026), and at chunk length one an adjacent-token switch penalty. The switch-only control we train is the symmetric total-variation form $1 - \sum_e \min(p_{t,e}, p_{t+1,e})$ on the full softmax, a mechanistic analogue of StickyMoE's published squared-ℓ_2 loss rather than a re-implementation (Proposition 3); the squared-ℓ_2 form is run as a further control. With residency frozen to a fixed set it becomes an offload penalty on off-resident mass (Proposition 4), the limit of MELINOE's cache-simulation loss when its decayed cache state is held fixed. For i.i.d. requests with stable marginals, caching the top-M blocks by marginal probability maximizes the expected hit rate among history-independent rules (Aho et al., 1971) (Lemma 2), the motivation for the memoryless form; the lemma assumes a fixed workload, which training moves.

4 Experimental design

Models and regimes. We pretrain OLMoE-architecture models (Muennighoff et al., 2025) on 2B FineWeb-Edu (Penedo et al., 2024) tokens. The primary geometry is 520M parameters (12 layers $\times$ 32 experts, top-4, 384 blocks), with a 64-expert control of near-equal parameter count (414M, 10 layers $\times$ 64 experts, top-8, 640 blocks) and a ladder from 62M to 6.8B (Section 5.4). Three regimes are compared. **Base** trains with λ=0. **Born** trains under Eq. 1 with λ ramping over the final 80% of training (about 1.6B pressured tokens). **Retrofit** continues training the finished base checkpoint under the same objective with all parameters trainable, a fresh AdamW state, 1% warm-up and cosine decay, and the λ ramp and τ anneal spread over its own budget; the *budget-matched* retrofit uses 1.6B tokens, the born arm's pressured budget, so both arms see the same schedules over the same tokens at $1.8\times$ the born arm's total compute (a born arm over 3.6B tokens matches it). Retrofits run at peak learning rates of 10^{-4} (a quarter of the pretraining peak, the customary choice for continued training), 2×10^{-4} and 4×10^{-4} (the pretraining peak), with a λ=0 continuation at 4×10^{-4} as the quality reference for the last. A released model, OLMoE-1B-7B, is adapted with only its router trainable (Section 5.4). Optimizer settings and every per-arm configuration are in Appendix A.

Capacities and the loss budget. Every memoryless pretraining arm trains with $M_\ell = 8$ per layer, which is 96 of 384 blocks at 32 experts (the 25% tier) and 80 of 640 at 64 experts (the 12.5% tier), except one control at $M_\ell = 16$; the born stateful arms use one sum constraint pooled over all layers' blocks, $\sum_{\ell,e} \rho_{\ell,e} = M$ with $M = 160$ (80 in one control), $W = 8$ and no switch term. Each arm records decode traces on WikiText-103 and we simulate LRU and Belady at 6.25, 12.5 and 25% of the blocks; the 25% tier (96 blocks at 32 experts, 160 at 64) is the reporting capacity unless stated, and the 64-expert results are shown at 12.5% (80 blocks) as well because the ordering among loss forms changes between the two.

Metrics. All traffic numbers are miss counts of a replacement policy simulated on a model's recorded routing traces at capacity M. LRU_M is the deployed policy; Belady_M is the offline optimum (Belady, 1966) and lower-bounds every online policy; Static_M is the miss count of a fixed pin of the M most frequently accessed blocks in each trace chunk, and $\kappa = \mathrm{LRU}_M/\mathrm{Static}_M$ is the temporal excess ($\kappa \geq 1$ is not a theorem; every arm in the main text has $\kappa \in [0.99, 1.9]$). LRU and Belady are simulated with mandatory insertion from a cold cache in each 1024-token chunk; the static pin serves non-resident accesses by bypass and pays no fill. A "multiple" is base misses divided by an arm's misses at the same capacity, seed means where seeds exist. Quality is WikiText-103 test perplexity (PPL), with a six-task zero-shot battery at every scale (Appendix C); exact definitions are in Appendix D.

Configuration differences among arms. Through an unscaled return value in the training script, every λ>0 pretraining and retrofit arm of the main sweep trained with the balance coefficient a at 1.0 rather than the intended library default 0.01, on every step, while the λ=0 base arms requested router logits every 50th step (effective $a \approx 0.02$); the error was found after the main sweep (Appendix A.5). Within the born family and between born and retrofit the coefficient is a constant; between the base reference and every λ>0 arm it is an uncontrolled factor, and since a base trained with the auxiliary at 1.0 on every step has 8–10% more LRU traffic, multiples against the plain base are conservative. We report the coefficient as an experimental variable with every arm's value in its table: the born-versus-retrofit comparison is replicated with $a = 0.01$ on every step in every arm (Section 5.3), the loss-form comparison is run at $a \in \{1.0, 0.3, 0.1, 0.01\}$ (Section 5.1), and the released-model

Table 1: Loss forms at 414M/64 experts (2B tokens, 1,966,080 accesses per trace): LRU and Belady misses at the 25% tier (160 of 640 blocks) and LRU at 12.5% (80); n seeds, seed means where $n > 1$; rows within a block ordered by perplexity. a is the balance coefficient (the base's auxiliary acted every 50th step, effective $a \approx 0.02$). Stateful arms pool one budget over all layers; every other arm uses M_ℓ=8 per layer except the M_ℓ=16 control. Per-seed values, Static and κ are in Appendix C.

arm	n	PPL	LRU_{160}	Belady_{160}	LRU_{80}
$a = 1.0$ *(the main runs)*					
base, λ=0	3	54.47	1,108,237	550,813	1,559,744
concentration-only λ=0.2	1	55.84	402,119	204,336	1,021,964
ours λ=0.1	2	58.34	228,614	115,098	931,492
switch-only λ=1.0	2	58.99	324,443	152,204	1,000,088
ours λ=0.2	3	61.16	168,166	83,188	763,651
switch-only λ=2.0	2	63.45	110,715	51,024	670,205
$a = 0.1$					
concentration-only λ=0.2	3	56.92	349,679	174,400	951,507
concentration-only λ=2.0	3	60.67	182,653	95,665	586,662
residency-only λ=0.2	3	63.48	59,623	31,493	653,391
ours λ=0.2	3	65.17	33,059	18,145	592,813
switch-only λ=2.0	3	65.41	72,176	33,349	565,054
residency-only λ=0.35	1	66.12	36,926	19,311	567,169
$a = 0.01$ *(library default)*					
memoryless, M_ℓ=16, λ=0.15	1	59.62	379,050	129,042	1,237,527
ours λ=0.1	1	61.75	62,607	32,042	697,310
squared-ℓ_2 switch-only λ=64	1	61.75	122,597	55,461	646,513
stateful λ=4.0, M=160	3	62.72	10,958	7,729	1,048,259
stateful λ=4.0, M=80	1	64.12	27,671	13,804	238,401
ours λ=0.15	3	65.11	27,544	15,223	611,124
switch-only λ=1.5	3	65.15	70,378	33,185	539,820
residency-only λ=0.25	3	66.05	20,876	12,093	586,735
switch-only λ=3.0	3	68.67	18,062	9,872	322,916
16 experts, λ=0 (160 of 160)	1	61.41	3,840	3,840	1,011,813

pipeline trained at 0.01 throughout. Run counts are given in every caption; a headline with one run per arm is labelled as such.

5 Results and analysis

5.1 Quality–traffic trade-offs among loss forms

Which objective buys the most traffic for a given perplexity, and at which capacity? Table 1 places our objective, its residency-only and stateful variants, the switch-only control and the hard-concentration control side by side at the 64-expert geometry at three balance coefficients, quality-matched pairs adjacent; the 32-expert counterparts are in Table 2 and Appendix C.

Pressure of any sufficient form moves the floor. At the coefficient of the main runs, concentration-only λ=0.2 reaches a floor of 204,336 at +1.4 PPL over base and switch-only λ=1.0 reaches 152,204 at +4.5, both below the 10^{-4} budget-matched retrofit's 214,654 at +5.5 (Table 2). The 16-expert model of the same active compute has only compulsory misses at 160 blocks, all of its blocks being resident, and 1,011,813 at 80; every ordering below is among 64-expert models and does not extend across expert counts.

Which form wins turns on the balance coefficient. At $a = 1.0$ the switch-only objective wins the heavy-pressure region at matched perplexity, by 1.55× fewer LRU misses at 32 experts on three seeds per arm (131,502 against 203,606 at 51.74 against 51.13 PPL) and by 1.24–1.35× at 64 experts on the two-seed interpolation between the λ=1.0 and 2.0 rows, and likewise for the pretraining-rate retrofit at 32 experts (Table 2); ours wins the mild region at 64 experts, where λ=0.1 moves 0.70× the misses of switch-only λ=1.0 at 0.65 PPL lower. Weakening the coefficient

reverses the heavy-pressure ordering. At $a = 0.1$, three seeds each, ours at 65.17 / 33,059 against switch-only at 65.41 / 72,176 moves $0.46\times$ the misses at the 25% tier and $1.05\times$ at 12.5%. At the library default the closest three-seed pair, ours λ=0.15 at 65.11 / 27,544 against switch-only λ=1.5 at 65.15 / 70,378 (means 0.04 ± 1.8 PPL apart; six-task battery 43.2 against 43.3, Table 8), gives $0.39\times$ at the 25% tier, below $0.5\times$ on all nine seed pairs, and $1.13\times$ at 12.5%. This pair had its reading rule fixed before the runs (Appendix A). The published squared-ℓ_2 switch form, run at four doses at $a = 0.01$, meets ours λ=0.1 at PPL 61.75 with $2.0\times$ the LRU misses at 25% and $0.93\times$ at 12.5%, one seed each.

The capacity dependence is the same at both weak coefficients. At the 12.5% tier, the tier at which the loss budget $M_\ell = 8$ was set, ours does not beat the switch-only control at matched perplexity, and switch-only λ=3.0 reaches 322,916 there at 3.6 PPL above the matched cell, $0.53\times$ our misses. The stateful form retrained against the 80-block budget lands at 64.12 / 238,401 at 12.5%, below every memoryless arm in the table at a lower perplexity than the matched cell, and ties the memoryless form at 25% (27,671 against 27,544). The favourable ordering for the residency objective is therefore a 25%-tier result at the budget used, and the stateful form is the one that carries to 12.5%.

5.2 Which components matter

Does the switch term carry the result, does soft residency add anything over hard top-M mass, and does stateful residency beat memoryless?

The residency term alone carries the trade-off. At $a = 0.01$ the residency-only arm at λ=0.25 lands at 66.05 / 20,876 on three seeds, 0.94 PPL above ours λ=0.15 with $0.76\times$ its LRU misses and $0.79\times$ its floor, within the 1-PPL window fixed in advance under which the switch term would be judged not to help materially; at $a = 0.1$ the residency-only arm at λ=0.35 sits 0.95 PPL above the full objective's three-seed mean with $1.12\times$ its LRU misses and $1.06\times$ its floor, the same Static and κ as the full objective's seed 0 (23,660 against 23,702; 1.56 against 1.57). At the same λ the residency-only arm is not quality-matched (at $a = 0.1$, 1.7 PPL below the full objective with $1.8\times$ its misses), so the switch term acts as extra dose more than as a distinct mechanism at these settings; a residual effect below $1.2\times$ is not excluded, since the one-seed residency-only arms were not placed below the full objective's perplexity.

Soft residency, not hard concentration. Replacing the soft residency by the hard top-M mass of the batch marginal (no τ) trades quality for traffic far less steeply. At $a = 0.1$ the hard form at λ=0.2 and 2.0 sits at 56.92 / 349,679 and 60.67 / 182,653 (three seeds each), no hard-form dose examined reaches the full objective's perplexity, and with the switch term added its LRU count stays $4.3\times$ the full objective's at λ=5 (Appendix C).

Stateful against memoryless. In born 64-expert training at $a = 0.01$ the stateful form at λ=4.0 lands at PPL 62.72 ± 1.30 with 10,958 LRU misses at the 25% tier on three seeds (per seed 61.22 / 11,243, 63.51 / 11,289, 63.42 / 10,341), below the memoryless matched cell on both axes, $0.40\times$ its misses at 2.4 PPL lower, with a floor of 7,729 against 15,223. Its $1.7\times$ deficit at 12.5% under the 160 budget is removed by training against 80 (Section 5.1); the memoryless form retrained at $M_\ell = 16$, the per-layer equivalent of the pooled 160, does not reproduce the stateful point (59.62 / 379,050). On the released model (Section 5.4) the two forms are indistinguishable on traffic gain at n=20 per arm ($+0.116 \pm 0.064$ against $+0.104 \pm 0.048$, Welch p=0.48), and the stateful form pays +0.06 more perplexity and violates the guard on 2 of 20 seeds against none. The stateful form is therefore the born recommendation and the memoryless form the retrofit one; neither dominates in both settings. Annealing τ has no measurable effect on the retrofit (n=10 per arm, p=0.88) and is kept because the limit $\tau \to 0$ is what Proposition 1 concerns.

Why the ordering reverses. Since $\mathrm{LRU} = \mathrm{Static} \times \kappa$, the quality-matched LRU ratio factors into a concentration ratio and a temporal ratio. At $a = 1.0$ every switch-trained arm has $\kappa \in [0.99, 1.09]$ over a $10\times$ range of λ and both geometries, and every marginal-trained arm, ours included, has $\kappa \in [1.22, 1.62]$; the switch term's $\approx 1.3\times$ advantage in κ exceeds our concentration lead there, and dose does not buy κ. Down the coefficient ladder the balance auxiliary opposes the miss term's Static directly (131k at 1.0 to 7.5k at 0.01 for ours λ=0.2) while switch-only λ=2.0's stays within 61–106k, which is why weakening it helps ours and not the control. An exact identity for κ (Appendix B, Eq. 3) locates the family separation in cold accesses recurring within four tokens (40–48% of cold

Table 2: Born against retrofit, misses at the 25% tier: 520M/32 experts (96 of 384 blocks, 1,179,648 accesses) and 414M/64 experts (160 of 640, 1,966,080). "LR" is the retrofit's peak learning rate (pretraining peak 4×10^{-4}), "tokens" the pressured budget for retrofits and the pretraining budget for born arms, a the balance coefficient of the block, n seeds per row (seed means where $n > 1$). The λ=0 continuation is the quality reference for the 4×10^{-4} retrofits. Dose steps, the onset sweep and per-capacity values are in Appendix C.

arm	LR	tokens	n	PPL	LRU	Belady
520M, 32 experts, $a = 1.0$						
base	—	2B	3	45.89	636,935	312,371
continued, λ=0	4×10^{-4}	1.6B	1	40.01	624,296	304,898
retrofit λ=0.2	10^{-4}	1.6B	2	50.20	351,866	139,491
retrofit λ=0.2	2×10^{-4}	1.6B	1	49.16	276,908	104,051
retrofit λ=0.2	4×10^{-4}	1.6B	1	45.16	182,960	65,699
born λ=0.1	—	2B	1	50.20	266,068	100,132
born λ=0.2	—	2B	3	51.13	203,606	70,852
born λ=0.2	—	3.6B	1	45.09	197,077	67,792
switch-only retrofit λ=1.0	4×10^{-4}	1.6B	1	44.50	143,565	67,795
switch-only born λ=1.0	—	2B	3	51.74	131,502	62,217
520M, 32 experts, $a = 0.01$ on every step in every arm						
base	—	2B	2	46.69	604,596	284,065
continued, λ=0	4×10^{-4}	1.6B	2	40.79	600,640	283,228
retrofit λ=0.2	4×10^{-4}	1.6B	2	44.18	174,639	62,423
born λ=0.2	—	2B	2	51.26	176,913	61,506
born λ=0.2	—	3.6B	1	43.33	145,862	51,431
414M, 64 experts, $a = 1.0$						
base	—	2B	3	54.47	1,108,237	550,813
retrofit λ=0.2	10^{-4}	1.6B	2	59.98	447,555	214,654
retrofit λ=0.2	4×10^{-4}	1.6B	1	53.01	152,762	75,162
born λ=0.1	—	2B	2	58.34	228,614	115,098
born λ=0.2	—	2B	3	61.16	168,166	83,188

accesses for switch-trained arms, 21–30% for residency-trained ones); a memoryless residency term cannot see how the cold residual is arranged in time, whereas the stateful residency moves with recency, consistent with its lead at 25%.

5.3 Pretraining under the objective versus retrofit

Must a model be pretrained under the objective to reach a given floor, or does continued training of a finished model reach the same point? Table 2 and Figure 1 give the comparison at both geometries and both coefficients.

Under the customary retrofit protocol the born model reaches a lower floor. With the retrofit's peak learning rate at 10^{-4}, the born λ=0.2 model's floor on three seeds is 70,852 against 139,491 for the retrofit given the same 1.6B pressured tokens (two seeds), a matched-budget gap of 1.97× at 0.9 PPL apart; a born arm at λ=0.1 lands at the retrofit's perplexity (50.20 against 50.20) with a floor 1.39× lower. The same protocol gives 2.58× at 64 experts (83,188 against 214,654) and 3.08× at 3.6B/64 experts, and the 64-expert retrofit's floor does not move at $a = 0.01$ (214,181; Appendix C).

The gap is set by the learning rate. A retrofit at the pretraining peak, 4×10^{-4}, with the same 1.6B tokens lands at PPL 45.16 with a floor of 65,699, 0.93× the born seed mean at 6.0 PPL lower. Its perplexity is below the base model's because continued training at the pretraining rate is itself more pretraining; against the λ=0 continuation with the same schedule (PPL 40.01 at the base model's traffic) it costs +12.9% PPL for a 4.64× lower floor, where the born arm costs +11.4% for 4.41×. At 64 experts the same retrofit reaches 75,162 at PPL 53.01, 0.90× the born seed mean at 8.2 PPL lower. The effect is monotone in the learning rate (139,491 → 104,051 → 65,699) and in tokens at the high rate (101,783 → 76,381 → 65,699 over 200M, 800M, 1.6B). The 3.6B-parameter retrofit ran only at 10^{-4}, so its 3.08× is a protocol number, and each 4×10^{-4} arm at $a = 1.0$ is one seed.

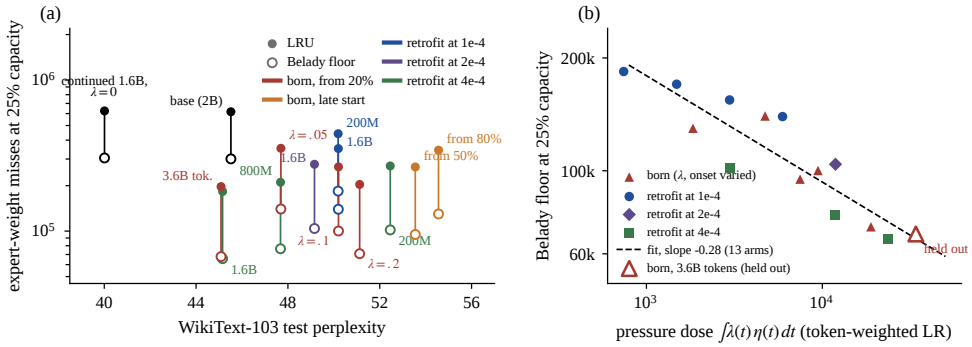

Figure 1: 520M/32 experts, $a = 1.0$, capacity 96 (25%). (a) Test perplexity against misses (filled, LRU; hollow, Belady floor; seed means where seeds exist): born arms (red), retrofits at 10^{-4} (blue) and 4×10^{-4} (green), the λ=0 continuation (black), born arms with onset at 50% or 80% (orange) and the 3.6B-token born arm. (b) Belady floor against the dose $D = \int \lambda(t)\,\eta(t)\,dt$ (η the learning-rate schedule) for thirteen arms with the 3.6B-token arm held out; one line fits born and retrofit arms (slope -0.28, r^2=0.78, born offset $0.98\times$).

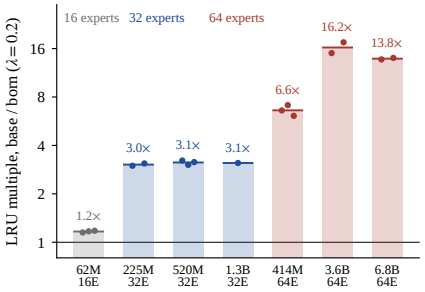

Figure 2: LRU traffic multiple (base over born, λ=0.2, 25% tier, $a = 1.0$) across seven pretraining geometries grouped by expert count; dots are per-seed multiples against the same-seed base, bars the seed mean (three seeds at 62M, 225M, 520M and 414M/64E; two at 3.6B and 6.8B; one at 1.3B). Per-arm values are in Table 7.

The closure replicates under a clean protocol. With $a = 0.01$ on every step in every arm, the 4×10^{-4} retrofit reaches a mean floor of 62,423 against 61,506 for the 2B-token born model, over two runs each ($1.02\times$; per seed 1.03 and 1.00), at 44.18 against 51.26 PPL, and 174,639 against 176,913 LRU misses. Against the two-seed λ=0 continuation the retrofit pays $+8.3\%$ PPL for a $4.54\times$ lower floor, the born arm $+9.8\%$ for $4.62\times$ against its base. The born operating point is therefore not a coefficient artefact ($3.42\times$ fewer LRU misses at $+4.6$ PPL at 0.01, against $3.13\times$ at $+5.2$ at 1.0), and neither is the closure.

Matched pressured tokens against matched total compute. At equal pressured budget the retrofit's total compute is $1.8\times$ the born arm's; a born arm over 3.6B tokens under the same ramp matches it. At $a = 1.0$ it ties the retrofit on both axes (45.09 PPL, floor 67,792, $1.03\times$ the retrofit's; LRU 197,077, $1.08\times$). At $a = 0.01$ it lands at 43.33 / 145,862 / 51,431, a floor $0.82\times$ the retrofit mean (0.87 and 0.79 against the two retrofit seeds) and $0.84\times$ its LRU misses at 0.85 PPL lower. The band fixed in advance for a tie was 0.85–$1.2\times$; this run falls outside it on the born side, so the tie prediction is reported as missed and the born model keeps a small floor lead at matched compute on one run against two (a second born seed is in progress).

Onset. Born arms that start the ramp at 50% and 80% of pretraining land at floors of 94,955 and 129,812 at PPL 53.55 and 54.56, above the 20% arm on both axes; a later start also means a lower learning rate at onset and a shorter ramp, so onset is not isolated. The dose D of Figure 1b fits born and retrofit arms with one line ($\log_{10}$ Belady $= 6.11 - 0.28 \log_{10} D$, born offset $0.98\times$) and places the held-out 3.6B-token arm at 65,885 against 67,792 observed, a descriptive fit at one geometry over mostly single-seed arms.

5.4 Scale and deployment relevance

Granularity, not parameter count, sets the multiple. Figure 2 repeats the protocol from 62M to 6.8B. The λ=0.2 LRU multiple is $1.15\times$ at 62M/16 experts, where two slots per layer cannot

hold top-4 routing and LRU saturates for every arm; 3.0–3.2× at 225M, 520M and 1.3B, all at 32 experts; and $6.6 \times \pm 0.5$ at 414M, 16.2× at 3.6B and 13.8× at 6.8B, the OLMoE geometry, all at 64 experts (two-seed means), at a relative perplexity cost of +7–13% throughout. Moving from 32 to 64 experts at near-constant parameters (520M to 414M) doubles the multiple; the further factor of two at 3.6–6.8B is confounded with depth and width. At 4× the training tokens the 520M multiple is unchanged and the 64-expert one grows to 7.9×; on out-of-domain C4 traces the 520M multiple is 3.32× against 3.22× in domain. Shuffling each trace preserves about 90% of the multiple, so the mechanism is marginal working-set concentration (286 of 384 experts stay in use at 520M).

Architectural controls. A 16-expert model of matched active compute at 520M costs +2.8 PPL and at equal fast-tier bytes (96 of its 192 blocks) moves 418,329 LRU misses with a floor of 165,566, so the born λ=0.05 arm (47.69 / 353,242 / 139,982) dominates it on both axes. Halving top-k (top-2 of 32) reaches 2.93× at +3.2 PPL with a floor 1.43× above the born arm's, and top-4 of 64 reaches only its 2.85× construction bound against 6.6×; against top-2 we claim parity, not superiority.

Downstream quality. On six zero-shot tasks the full-pressure born arm costs at most about one mean accuracy point at every scale (Table 8; one run per arm, several tasks near chance at 414M, so this bounds a large regression only), and the adapted released model is within 0.1 points of its base (70.79 against 70.87).

A released model. Fine-tuning only the router of OLMoE-1B-7B on 2M WikiText-103 tokens under the memoryless objective, with fresh decode traces per candidate and a guard of ΔPPL ≤ 0.3 against the λ=0 fine-tuning floor taken over seeds, the miss-only term at λ=0.15 reduces cache misses by $+0.116 \pm 0.064$ over twenty seeds with no guard violation, an expected gain of about 11% (a recipe selected by automated search does not hold on refresh; Appendix A.2). Under the same guard, model, data and budget the switch-only control tuned over a 20× dose range plateaus at +0.104 to +0.105 (ten seeds each) and exceeds ours only where every seed violates the guard, so neither form's advantage is demonstrated at these seed counts (Welch $p \geq 0.34$; Table 9); the released-model result does not show superiority over the switch-only baseline. Crossing {base, adapted router} × {LRU, evolved cache policy} is super-additive only at the 6% tier (synergy +13.0%, 95% interval excluding zero) and additive at 12.5 and 25%.

Wall clock. Replaying recorded traces through a two-tier H100 harness (Appendix D; missed blocks copied before the forward pass, "blocking", or on a concurrent stream, "overlapped", an upper bound) converts traffic to time. The replay matches the simulator's byte count for the released model (1.613 against 1.61 GB per token), whose adapted router moves 0.63 against 0.80 GB per token at the 25% tier for +16.9% overlapped and +14.4% blocking decode throughput. For the 6.8B born replica the 12.5% tier's traffic falls 1.056 → 0.425 GB per token for +63% blocking and +90% overlapped; at 25% decode becomes compute-bound and the overlapped gain saturates at +15%, and at 6.25% both arms thrash (Table 11). The 13.8× and the +90% are at different tiers and are not one conversion; every replay is a single run and none pays the expert-parallel cost of concentrated routing.

6 Discussion and limitations

Three results stand, each with a stated limit. The traffic reduction is set by expert granularity and converts to measured decode time, but the 3.6B and 6.8B models saw 0.3 tokens per parameter, depth is a confound we did not remove, quality above 520M is perplexity plus a single-run battery, and the replays are single runs. The residency objective wins at the 25% tier and the switch-only control at 12.5% once the balance coefficient is at or below 0.1, while at the coefficient of the main runs the control wins heavy pressure; the ordering rests on three seeds per arm at one 64-expert geometry, the control is an analogue of StickyMoE whose published form entered no matched window on one seed per dose, and on the released model neither form's advantage is demonstrated. The born–retrofit gap is a learning-rate effect and closes under a clean protocol on two seeds per arm; at matched total compute the protocols tie at one coefficient and the born model leads by 1.2× at the other on one run, so whether an onset effect survives at the library default is open. We recommend the stateful form for pretraining and the memoryless form for router-only adaptation, with the budget set to the deployment tier.

USE OF LARGE LANGUAGE MODELS

Automated search driven by an LLM was used to select two artifacts evaluated in this paper, a router fine-tuning recipe and a cache policy, each then measured under the same guards as its hand-written counterpart; the procedure and its archives are released. LLMs were also used as a writing aid. The claims, experimental design and analyses are the authors'.

REPRODUCIBILITY STATEMENT

All recipes, traces, fidelity records and scripts are released and every number in the paper regenerates from the accompanying repository. Appendix A gives every training setting and the selection and replication procedures, Appendix C the per-run values behind each table and figure, and Appendix D the trace, simulator and replay conventions and the artifact guide.

REFERENCES

Alfred V. Aho, Peter J. Denning, and Jeffrey D. Ullman. Principles of optimal page replacement. *Journal of the ACM*, 18(1):80–93, 1971.

Andrii Balashov and Olena Ponomarova. TriRoute: Unified learned routing for joint adaptive attention, experts, and KV-cache allocation. *arXiv preprint arXiv:2607.06601*, 2026.

L. A. Belady. A study of replacement algorithms for a virtual-storage computer. *IBM Systems Journal*, 5(2):78–101, 1966.

Andrew Blake and Andrew Zisserman. *Visual Reconstruction*. MIT Press, 1987.

Shaoke Fang, Ziang Li, Wenfei Wu, Jiatong Ji, Qingsong Liu, and Ruizhi Pu. Not all tokens are worth caching: Learning semantic-aware eviction for LLM prefix caches. *arXiv preprint arXiv:2605.18825*, 2026.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Elad Hazan, Kfir Yehuda Levy, and Shai Shalev-Shwartz. On graduated optimization for stochastic non-convex problems. In *International Conference on Machine Learning (ICML)*. PMLR, 2016.

Duc Hoang, Ajay Jaiswal, Mohammad Samragh, and Minsik Cho. SpecMD: A comprehensive study on speculative expert prefetching. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.03921.

Zachary Izzo, Lexing Ying, and James Zou. How to learn when data reacts to your model: Performative gradient descent. In *International Conference on Machine Learning (ICML)*. PMLR, 2021.

Ali Kayyam. Sticky routing: Training MoE models for memory-efficient inference. *arXiv preprint arXiv:2607.08780*, 2026.

John P. Miller, Juan C. Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning (ICML)*. PMLR, 2021.

Luca Moschella, Laura Manduchi, and Ozan Sener. Learning to evict from key-value cache. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.10238.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoE: Open mixture-of-experts language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2409.02060.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning (ICML)*, volume 119. PMLR, 2020.

Arian Raje, Anupam Nayak, and Gauri Joshi. MELINOE: Fine-tuning enables memory-efficient inference for mixture-of-experts models. *arXiv preprint arXiv:2602.11192*, 2026.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Andrii Skliar, Ties van Rozendaal, Romain Lepert, Todor Boinovski, Mart van Baalen, Markus Nagel, Paul Whatmough, and Babak Ehteshami Bejnordi. Mixture of cache-conditional experts for efficient mobile device inference. *Transactions on Machine Learning Research*, 2025. arXiv:2412.00099.

Chenyang Song, Weilin Zhao, Xu Han, Chaojun Xiao, Yingfa Chen, Yuxuan Li, Zhiyuan Liu, and Maosong Sun. BlockFFN: Towards end-side acceleration-friendly mixture-of-experts with chunk-level activation sparsity. *arXiv preprint arXiv:2507.08771*, 2025.

An Wang, Xingwu Sun, Ruobing Xie, Shuaipeng Li, Jiaqi Zhu, Zhen Yang, Pinxue Zhao, J. N. Han, Zhanhui Kang, Di Wang, Naoaki Okazaki, and Cheng-zhong Xu. HMoE: Heterogeneous mixture of experts for language modeling. *arXiv preprint arXiv:2408.10681*, 2024.

Ziteng Wang, Jun Zhu, and Jianfei Chen. ReMoE: Fully differentiable mixture-of-experts with ReLU routing. In *International Conference on Learning Representations (ICLR)*, 2025.

Conventions. Run names in the released repository follow *codesign* for born arms, *posthoc* for retrofits, *g64* for the 414M/64-expert geometry and *sdn* for seed n. "Balance coefficient 1.0" is the unscaled value at which the main runs trained (Appendix A.5); "0.01" is the library default. Every table states its run count; a value with $n > 1$ is a seed mean, and per-seed values are listed wherever a headline rests on them.

A Implementation and experimental protocol

A.1 Pseudocode of both forms with their stop-gradients

The training step of Section 3.1 as implemented (`soft_miss_and_switch` in `experiments/04_pretrain_codesign/pretrain_moe.py`). $\mathrm{sg}[\cdot]$ is a stop-gradient; every quantity computed inside it is a constant to the backward pass.

for each layer ℓ with router logits $z \in \mathbb{R}^{T\times E}$:
  $p \leftarrow \mathrm{softmax}(z)$ (full softmax, not the top-k gate)
  memoryless: $u \leftarrow \frac{1}{T}\sum_t p_t$ (batch marginal, one vector)
  stateful: $u_t \leftarrow \sum_{k\geq 1} \frac{1}{W}(1-\frac{1}{W})^{k-1} p_{t-k}$, $u_1 \leftarrow 0$
    (causal EMA of strict-past gates, one vector per token)
  $\tau' \leftarrow \mathrm{sg}[\max(\tau(t)\cdot \mathrm{std}(u),\, 10^{-6})]$
  $b \leftarrow \mathrm{sg}[\,\mathrm{bisection}_{40}\{b : \sum_e \sigma((u_e - b)/\tau') = M\}\,]$
    (per token in the stateful form)
  $\rho \leftarrow \sigma((u-b)/\tau')$ (gradient flows through u only)
  memoryless: $\mathcal{L}^{\ell}_{\mathrm{miss}} \leftarrow \sum_e u_e(1-\rho_e)$
  stateful: $\mathcal{L}^{\ell}_{\mathrm{miss}} \leftarrow \frac{1}{T}\sum_t \sum_e p_{t,e}(1-\rho_{t,e})$
  $\mathcal{L}^{\ell}_{\mathrm{sw}} \leftarrow 1 - \frac{1}{T-1}\sum_t \sum_e \min(p_{t,e}, p_{t+1,e})$
$\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{LM}} + a\,\mathcal{L}_{\mathrm{bal}} + \lambda(t)\,\big[\, w_{\mathrm{m}}\,\mathrm{mean}_{\ell}\mathcal{L}^{\ell}_{\mathrm{miss}}$
    $+\, w_{\mathrm{s}}\,\mathrm{mean}_{\ell}\mathcal{L}^{\ell}_{\mathrm{sw}}\,\big]$
backward on $\mathcal{L}$; AdamW step

Arms labelled "ours" use $(w_{\mathrm{m}}, w_{\mathrm{s}}) = (1, 0.5)$, residency-only $(1, 0)$, switch-only $(0, 1)$. In the memoryless form $M = M_\ell$ per layer; the born stateful arms take the top-M once over the concatenated $[T, LE]$ gate matrix with M=160 (a control at 80), W=8 and w_{s}=0, whereas the released-model stateful arms use W=64 and a per-layer budget. The backward pass is the derivative of the forward value with b and τ' frozen; the omitted implicit term is given in Section 3.1.

Baseline implementations. The switch-only control replaces the bracket by $\mathcal{L}_{\mathrm{sw}}$ alone (the total-variation form above); the squared-ℓ_2 control uses $\frac{1}{T-1}\sum_t \|p_t - p_{t+1}\|_2^2$, the published StickyMoE form, and needs doses of $\lambda \in \{2, 4, 16, 64\}$ to move traffic. The concentration-only control replaces $\mathcal{L}^{\ell}_{\mathrm{miss}}$ by $1 - \sum_{e \in \text{top-}M_\ell(u)} u_e$ with no τ, and "concentration + switch" adds the half-weight switch term to it. The BlockFFN analogue penalizes routed mass on experts unused in the previous chunk of 8 tokens (Proposition 3 with $L = 8$). Every control shares the λ ramp, the τ anneal where it applies, the optimizer and the data order of the arm it is compared with.

A.2 Training settings

Pretraining. OLMoE-architecture models (HF `OlmoeForCausalLM`) on 2B FineWeb-Edu tokens with AdamW ($\beta = (0.9, 0.95)$, weight decay 0.1, gradient clip 1.0), peak learning rate 4×10^{-4}, 1% warm-up, cosine decay, sequence length 1024, 128 sequences per step, bf16. Geometries (layers × experts, top-k, blocks): 62M (8×16, top-4, 128); 225M (10×32, top-4, 320); 520M (12×32, top-4, 384); 414M (10×64, top-8, 640); 1.3B (16×32, top-4, 512); 3.58B (18×64, top-8, 1152); 6.8B (16×64, top-8, 1024, the OLMoE-1B-7B geometry). Token-axis controls at 520M and 414M/64E train on 8B tokens. The λ ramp is linear from zero at 20% of training to its plateau at the final step; τ anneals geometrically from 0.5 to 0.05 over the same span. Per-layer budget $M_\ell = 8$ throughout, except the $M_\ell = 16$ control and the pooled stateful budgets above.

Retrofit. Continued training of the finished base checkpoint with all parameters trainable, a fresh AdamW state, the same hyper-parameters, 1% warm-up and cosine decay over the retrofit budget, and the λ ramp and τ anneal from the first step spread over that budget. Budgets are 200M tokens (400M and 800M in the dose steps) and 1.6B (budget-matched); peak learning rates 10^{-4}, 2×10^{-4} and 4×10^{-4}. The λ=0 continuation runs the 4×10^{-4} retrofit schedule with the traffic term off.

Released-model adaptation. OLMoE-1B-7B with only the router weights trainable, fine-tuned on 2M WikiText-103 training tokens under the memoryless objective with the same anneal, $M_\ell = 8$ (4 and 16 in the cross-budget sweep), and a quality guard of ΔPPL ≤ 0.3 against the λ=0 fine-tuning floor (in-domain fine-tuning alone improves PPL 9.179 → 8.373, so quality is charged against the floor), taken as the maximum over seeds. Each candidate records fresh decode traces (24 chunks of 1024 tokens, 128 accesses per token) and is scored by the reduction in the summed miss count of a fixed cache policy at capacities $\{64, 128, 256\}$ against the un-adapted model's own traces (`experiments/03_coadapt_real/worker_ft.py`). The evolved cache policy of Table 10 came from two hundred generations of automated search on the base model's traces with disjoint train and hold-out chunks (score 0.392 train, 0.386 hold-out, as fraction of the LRU-to-Belady gap closed); the router recipe "evo009" came from the same search procedure on the objective's hyper-parameters, selected on a single seed, which is why its selection draw (+21.9%) exceeds its nine-seed refresh ($+0.111 \pm 0.033$) and why the plain miss-only term at λ=0.15 over twenty seeds is the number reported.

A.3 Selection, replication and run-count conventions

Every headline is a seed mean over the runs stated in its table or caption, with the single-run cases labelled. Quality-matched pairs are the two arms with the closest mean perplexities among those run; where no pair falls within 1 PPL, no ratio is claimed. Multiples are base misses over arm misses at the same capacity and policy, the base being the same-geometry, same-budget arm as trained (auxiliary on every 50th step); a base retrained with the auxiliary on every step moves +8.0% LRU at 64 experts (three seeds, PPL 54.40 against 54.47) and +10.2% at 520M (two seeds, 46.14 against 45.52), so the multiples quoted are the conservative ones, and because the 520M duty-matched base is +0.62 PPL worse its extra traffic is not attributed. Regenerating a released-model recipe with an identical seed moved its simulated gain from +0.32 to +0.10 while the λ=0 control reproduced within 0.7%; traffic pressure amplifies bf16 nondeterminism into distinct routing solutions, so run-

to-run spread is reported as mean ± std over seeds and hardware replays are compared with the replayed artifact's own simulated gain, never with a fleet mean.

A.4 Which comparisons were fixed in advance

The study was exploratory in its first phase and criteria-fixed in its second. The following comparisons had their predictions, seed counts and reading rules written before the runs, and the outcomes reported are those rules applied as written: the 520M born-versus-retrofit comparison under the 10^{-4} protocol (met) and its 4×10^{-4} repeat with the rule that a retrofit floor within $1.1\times$ of the born line at its perplexity withdraws the born advantage (it did); the 3.6B-token born arm at $a = 1.0$ with a tie band of 0.85–$1.2\times$ (met); the clean-coefficient replication at 0.01 with the falsifiers "retrofit floor $\geq 1.3\times$ born" and "retrofit $\leq 0.7\times$" (neither fired) and its 3.6B-token born arm with the same tie band (missed, born side); the coefficient ladder at 0.1 and 0.01 with three seeds per cell and the quality-overlap fill (ours λ=0.15, switch-only λ=3.0) with the rule that the comparison cell is the pair of three-seed arms with the closest mean perplexities and the claim requires $\leq 0.7\times$ within 1.0 PPL on every seed pair (met at $0.39\times$); the residency-only arms at both weak coefficients with the 1-PPL, $1.2\times$ rule (Section 5.2); the stateful born arm and its 80-budget control; the duty-matched base with the falsifier that it moves less traffic than the plain base (did not fire); the squared-ℓ_2 switch-only doses; and the released-model switch-only fills at $\lambda \in \{1.25, 1.5\}$ (λ=1.5 was predicted to violate the guard on four to ten seeds and did not). The dose sweeps, the onset arms, the scaling ladder, the architectural controls, the hard concentration arms and the released-model recipe search were exploratory, and the dose fit of Figure 1b was made after those arms with one held-out point. Three further variants of the objective (a recency-weighted residency miss term, a switch term restricted to non-resident blocks, and a miss $\times$ switch grid) were run under stated criteria, failed them, and are recorded in the repository log (`docs/OBJECTIVE_V2.md`) rather than here; the same log holds the original prediction records.

A.5 The balance-coefficient protocol note

The pretraining script passes no labels to the HF OLMoE forward and adds the returned `aux_loss` to the LM loss; in that path the library returns the Switch-style balance loss unscaled, so the configured `router_aux_loss_coef`=0.01 was never applied in the main sweep. Every λ>0 pretraining and retrofit arm of Tables 1 and 2 marked $a = 1.0$ therefore trained with the balance loss at coefficient 1.0, on every step; the λ=0 base arms requested router logits every 50th step and trained at an effective ≈0.02. The arms marked $a \in \{0.3, 0.1, 0.01\}$ apply the stated coefficient on every step. The error was found after the main sweep and before the ladder and the clean replication, which were then registered as above.

B Supporting mathematical statements

Notation: a trace of T tokens, token t accessing the set S_t of A blocks; $N = TA$ accesses; capacity M; f_e the access count of block e and $\bar{a} = f/N$.

Proof of Proposition 1. For any $\rho \in [0,1]^E$ with $\sum_e \rho_e = M$, $\sum_e \bar{a}_e \rho_e \leq \mathrm{Top}_M(\bar{a})$: the objective is linear in ρ over the polytope $\{0 \leq \rho \leq 1,\ \sum \rho = M\}$, whose vertices are the hard M-subset indicators, so the maximum is attained at the top-M set. The soft residency of Eq. 2 satisfies the constraint at every τ, giving $\sum_e \bar{a}_e(1-\rho_e) \geq 1 - \mathrm{Top}_M(\bar{a})$; without the constraint the bound fails (take $\rho \equiv 1$). Next, $N(1 - \mathrm{Top}_M(\bar{a}))$ is the number of accesses to blocks outside the top-M set, which is the miss count Static_M of pinning that set and serving every other access by bypass. Charging the pin its M first-touch fills makes it a policy in the class that may serve a miss with or without insertion, whose offline optimum satisfies $\mathrm{Belady}_M^{\mathrm{byp}} \leq \mathrm{Static}_M + M$. The demand-paging Belady of the tables is not bounded by a bypass policy in general (capacity one, trace a, b, a, b: demand Belady misses four, the pin of a misses two plus one fill); on all 136 traces of this paper $\mathrm{Belady}_M \leq \mathrm{Static}_M + (\text{chunks})\, M$ holds with each chunk's cold start charged as a fill. □

Lemma 1 (window-union hit condition)**.** *Under LRU with capacity M, an access at token $t{+}1$ to a block last accessed at token $t{-}w{+}1$ or later is a hit whenever $|S_{t-w+1} \cup \cdots \cup S_{t+1}| \leq M$.*

Proof. The distinct blocks touched since that block's last access are a subset of the union minus the block itself, so fewer than M distinct others intervened and LRU has not evicted it. □

Corollary 1. *(a) If $M \geq 2A$ then $\mathrm{Belady}_M \leq \mathrm{LRU}_M \leq A + \mathrm{Sw}_1$ with $\mathrm{Sw}_1 = \sum_t |S_{t+1} \setminus S_t|$. (b) If each token accesses k distinct blocks in each of L layers in a fixed layer order, $A = Lk$, and $M \leq A - k$, then $\mathrm{LRU}_M = N$ for every trace.*

Proof. (a) $|S_t \cup S_{t+1}| \leq 2A \leq M$, so by Lemma 1 with $w = 1$ every access in $S_t \cap S_{t+1}$ hits; the misses at token $t{+}1$ are at most $|S_{t+1} \setminus S_t|$, and the first token misses at most A. (b) Between two consecutive accesses to a block of layer ℓ the trace touches the k blocks of every other layer, $(L-1)k = A-k \geq M$ distinct others, so LRU has evicted it before its reuse; the $k-1$ other blocks of the same layer need not be distinct across tokens, which is why the condition is $M \leq A - k$ and not $M < A$. (At capacity $40 \leq A - k = 72$ every 64-expert arm records exactly 1,966,080 LRU misses, and at $40 \leq A - k = 44$ every 520M arm 1,179,648.) □

For $A \leq M < 2A$, (a) holds under the per-token condition $|S_{t+1} \setminus S_t| \leq M - A$ and can fail otherwise. Neither Static_M nor Sw_1 bounds the other in general: cyclic access to $M{+}1$ blocks gives $\mathrm{LRU}_M = N$ against $\mathrm{Static}_M \approx N/(M{+}1)$, while grouped access (each block's accesses contiguous) gives $\mathrm{LRU}_M = \#\text{distinct} \leq \mathrm{Static}_M$. Since $\bar{a}$ is invariant to token permutations, no function of the marginal alone controls κ.

Lemma 2 (marginal residency is optimal for i.i.d. requests)**.** *For i.i.d. requests with marginals $\bar{a}$ over unit-size blocks, a history-independent randomized residency rule is a distribution over subsets with per-block inclusion probabilities $q \in [0,1]^E$, $\sum_e q_e \leq M$, chosen independently of the requests. Its expected hit rate is $\sum_e \bar{a}_e q_e$, maximized by $q = \mathbf{1}\{e \in \text{top-}M(\bar{a})\}$.*

Proof. By independence of rule and request the hit probability at any step is $\sum_e \bar{a}_e q_e$, linear in q over the polytope $\{q \in [0,1]^E : \sum_e q_e \leq M\}$; a linear function attains its maximum at a vertex, and the maximizing vertices are indicators of M largest-$\bar{a}$ blocks (Aho et al., 1971). □

Proposition 2 (static-placement limit)**.** *Under i.i.d. accesses with marginals $\bar{a}$ and fixed $\tau > 0$, the stateful objective converges as $W \to \infty$ (with $t/W \to \infty$) to $\mathbb{E}_x\big[\sum_e p_\theta(e \mid x)\, w_e\big]$ with fixed weights $w_e = s_e c_{\mathrm{miss}}\big(1 - \rho_e(\bar{a})\big)$, a static tier-placement penalty on non-resident routing mass (Skliar et al., 2025; Wang et al., 2024).*

Proposition 3 (chunk-consistency limit)**.** *With hard residency ($\tau \to 0$) taken as the union U of blocks accessed in the previous length-L chunk, which requires relaxing the sum constraint to $\sum_e \rho_e = |U|$, the objective equals c_{miss} times the expected routed mass on blocks unused in that chunk, a chunk-sticky consistency objective (Song et al., 2025; Kayyam, 2026); for $L = 1$ it is an asymmetric adjacent-token switch penalty.*

Proposition 4 (single-boundary limit)**.** *With residency frozen to a fixed set R, $|R| = M$, and constant miss cost across one fast/slow boundary, the objective is proportional to the off-resident routed mass $\mathbb{E}_x\big[\sum_{e \notin R} p_\theta(e \mid x) s_e\big]$ (Raje et al., 2026; Wang et al., 2025).*

Throughout, fix θ and write the per-token traffic term of Eq. 1 as $T(u) = \sum_e p_\theta(e \mid x)\, s_e\, (1 - \rho_e(u))\, c_{\mathrm{miss}}$.

Proof of Proposition 2. Let accesses $a_t \in \{0,1\}^E$ be i.i.d. with $\mathbb{E}[a_{t,e}] = \bar{a}_e$, and let u_t be the EMA with rate $\beta = 1/W$. Then $\mathbb{E}[u_{t,e}] = \bar{a}_e + (1-\beta)^t(u_{0,e} - \bar{a}_e)$ and, by independence, $\mathrm{Var}(u_{t,e}) \leq \frac{\beta}{2-\beta}\,\bar{a}_e(1-\bar{a}_e)$. As $W \to \infty$ with $t/W \to \infty$ both vanish, so $u_t \to \bar{a}$ in L^2. The sum-constrained map $u \mapsto \rho(u)$ is continuous for $\tau > 0$ with the scale floored away from zero: at fixed scale the threshold $b_\tau(u)$ is coordinate-wise monotone and translation-equivariant, hence 1-Lipschitz in $\|\cdot\|_\infty$; $\sigma(\cdot/\tau')$ is $(4\tau')^{-1}$-Lipschitz; and $u \mapsto \mathrm{std}(u)$ is continuous. Since T is a bounded continuous function of ρ, $\mathbb{E}[T(u_t)] \to T(\bar{a})$. □

Proof of Proposition 3. Partition tokens into chunks of length L and set the hard residency during chunk C_j to $\rho_e = \mathbf{1}\{e \in U_{j-1}\}$, U_{j-1} the set of blocks accessed in C_{j-1} (well defined when $M \geq |U_{j-1}|$). Substituting into T gives, for $x \in C_j$, $T = c_{\mathrm{miss}} \sum_{e \notin U_{j-1}} p_\theta(e \mid x)\, s_e$. For $L = 1$, U_{j-1} is the previous token's access set. □

Proof of Proposition 4. With $\rho = \mathbf{1}_R$ and $c_{\text{miss}} \equiv c$, substitution gives $T = c \sum_{e \notin R} p_\theta(e \mid x)\, s_e$. □

The κ identity. Fix a chunk and call its top-M blocks hot, the rest cold, so cold accesses number Static_M. Under LRU from a cold cache, hot misses equal hot evictions plus the hot blocks resident when the chunk ends, H_{end} (summed over the C chunks; $H_{\text{end}} \le MC$), and evictions equal misses less the MC fills. With h_c the cold hit fraction and q the hot-victim fraction of evictions, $\text{LRU} = \text{Static}(1 - h_c) + q\,(\text{LRU} - MC) + H_{\text{end}}$, so exactly

$$\kappa \;=\; \frac{(1 - h_c) + (H_{\text{end}} - qMC)/\text{Static}}{1 - q}, \tag{3}$$

which reproduces the measured κ to three decimals on every trace (`scripts/kappa_decomp.py`). The boundary term is $O(MC/\text{Static})$; dropping it gives $(1 - h_c)/(1 - q)$, within 0.03 of κ where Static exceeds 100k ($MC = 3840$) and off by up to 0.46 for the most concentrated arms. Two consequences. $\kappa \ge 1$ is not a theorem (grouped access gives $\kappa < 1$, and one switch-only seed has 0.99); the floor is $\kappa \ge \text{Belady}/\text{Static}$, so on a fixed trace a policy change can gain at most $\kappa/(\text{Belady}/\text{Static}) \approx 1.32/0.62 \approx 2.1\times$ for our arms, a bound on the policy lever and not on the trainable one. The converse, that a Static lever is unbounded at equal κ, is false: as $\text{Static} \to 0$ the working set approaches M, $q \to 1$ and κ diverges. Table 6 lists the terms for the arms of Section 5.2.

C Detailed experimental results

C.1 520M/32 experts, all arms and capacities

Table 3 lists every 520M arm at $a = 1.0$ at both non-saturated capacities (at 24 blocks, 6.25%, every arm saturates LRU by Corollary 1b), and Table 4 the per-seed values of the clean-coefficient replication.

Table 3: 520M/32 experts, $a = 1.0$, 2B-token protocol; misses over 1,179,648 accesses, percentages against base. Seed 0 per arm; base and born λ=0.2 have three seeds and the 10^{-4} 1.6B retrofit two (seed means in Table 2), every other row is one run. The switch-only born λ=1.0 row is a three-seed mean.

arm	PPL	LRU@48	LRU@96	Belady@48	Belady@96
base	45.52	878,456	616,637	525,293	300,609
retrofit λ=0.05, 10^{-4}, 200M	46.08 (+0.6)	—	572,024 (−7%)	—	264,095 (−12%)
retrofit λ=0.2, 10^{-4}, 200M	50.19 (+4.7)	802,269 (−9%)	441,202 (−28%)	434,367 (−17%)	184,003 (−39%)
retrofit λ=0.2, 10^{-4}, 1.6B	50.21 (+4.7)	770,209 (−12%)	353,797 (−43%)	398,301 (−24%)	140,499 (−53%)
retrofit λ=0.2, 10^{-4}, 1.6B, seed 1	50.18 (+4.7)	768,257 (−13%)	349,934 (−43%)	396,405 (−25%)	138,482 (−54%)
retrofit λ=0.5, 10^{-4}, 1.6B	57.70 (+12.2)	706,857 (−20%)	210,021 (−66%)	339,570 (−35%)	73,861 (−75%)
retrofit λ=0.2, 2×10^{-4}, 1.6B	49.16 (+3.6)	736,242 (−16%)	276,908 (−55%)	364,339 (−31%)	104,051 (−65%)
retrofit λ=0.2, 4×10^{-4}, 200M	52.46 (+6.9)	729,556 (−17%)	269,940 (−56%)	361,578 (−31%)	101,783 (−66%)
retrofit λ=0.2, 4×10^{-4}, 800M	47.68 (+2.2)	691,184 (−21%)	210,628 (−66%)	331,627 (−37%)	76,381 (−75%)
retrofit λ=0.2, 4×10^{-4}, 1.6B	45.16 (−0.4)	653,587 (−26%)	182,960 (−70%)	313,986 (−40%)	65,699 (−78%)
continued λ=0, 4×10^{-4}, 1.6B	40.01 (−5.5)	884,233 (+1%)	624,296 (+1%)	530,494 (+1%)	304,898 (+1%)
born λ=0.05	47.69 (+2.2)	—	353,242 (−43%)	—	139,982 (−53%)
born λ=0.1	50.20 (+4.7)	724,890 (−17%)	266,068 (−57%)	359,163 (−32%)	100,132 (−67%)
born λ=0.2	51.35 (+5.8)	676,600 (−23%)	191,580 (−69%)	322,752 (−39%)	66,640 (−78%)
born λ=0.2, onset 50%	53.55 (+8.0)	739,080 (−16%)	265,819 (−57%)	363,797 (−31%)	94,955 (−68%)
born λ=0.2, onset 80%	54.56 (+9.0)	770,940 (−12%)	343,445 (−44%)	400,342 (−24%)	129,812 (−57%)
born λ=0.2, 3.6B tokens	45.09 (−0.4)	692,447 (−21%)	197,077 (−68%)	329,165 (−37%)	67,792 (−77%)
switch-only born λ=0.2	46.51 (+1.0)	684,343 (−22%)	341,496 (−45%)	363,200 (−31%)	168,430 (−44%)
switch-only born λ=1.0 (3-seed mean)	51.74 (+6.2)	450,225 (−49%)	131,502 (−79%)	222,384 (−58%)	62,217 (−79%)
switch-only retrofit λ=1.0, 4×10^{-4}, 1.6B	44.50 (−1.0)	477,714 (−46%)	143,565 (−77%)	236,475 (−55%)	67,795 (−77%)
16 experts, λ=0 (192 blocks)	48.67 (+3.2)	789,962 (−10%)	418,329 (−32%)	424,800 (−19%)	165,566 (−45%)
top-2 of 32, λ=0	48.74 (+3.2)	—	210,615 (−66%)	—	94,995 (−68%)

Table 4: Clean-coefficient replication at 520M/32 experts, $a = 0.01$ on every step in every arm, capacity 96; per-seed values behind the $a = 0.01$ block of Table 2. The retrofit of each seed starts from the base of the same seed. The 16-expert control at capacity 48, 96 and 144 of its 192 blocks moves 789,962, 418,329 and 138,605 LRU misses (Belady 424,800, 165,566, 47,938).

arm	PPL (sd0 / sd1)	LRU@96 (sd0 / sd1)	Belady@96 (sd0 / sd1)
base	46.78 / 46.59	594,317 / 614,875	277,704 / 290,426
continued λ=0, 4×10^{-4}, 1.6B	40.78 / 40.79	598,535 / 602,745	281,281 / 285,174
retrofit λ=0.2, 4×10^{-4}, 1.6B	44.26 / 44.09	166,174 / 183,104	59,387 / 65,458
born λ=0.2, 2B	50.86 / 51.65	165,842 / 187,984	57,548 / 65,463
born λ=0.2, 3.6B	43.33 / —	145,862 / —	51,431 / —

C.2 414M/64 EXPERTS, PER-SEED VALUES

Table 5 gives every seed behind Table 1 together with the dose sweeps at $a = 1.0$, the $a = 0.3$ rung, the concentration + switch arms and the single-run points at 0.01. Eleven further 64-expert arms at $a = 1.0$ that combined the miss and switch terms at other weights, gated one term by the other, or used a per-sequence marginal, all of which trained to perplexities of 63.6–73.4 without improving on the arms below at matched quality, and the released-model variants that failed their stated criteria, are listed with their values in the repository inventory (`data/arm_inventory.csv`) and not here.

Re-simulating the 64-expert traces at $M \in \{A, 2A, 3A, 4A\}$ (`scripts/kappa_vs_cap.py`) brackets where the switch term's advantage at $a = 1.0$ lives: at $M = A$, where Corollary 1(a) does not apply, the quality-matched heavy-pressure pair is at parity (ratio 1.03); at $2A$ and $3A$ the switch-only arm wins it by $1.38\times$ and $1.61\times$; at $4A$ both traces nearly fit and the distinction collapses. The mild-pressure pair favours ours at every capacity (ratio 0.66–0.89).

Table 5: 414M/64 experts, 2B tokens, capacity 160 (25%); one entry per seed, "PPL / LRU / Belady". Static_{160} and κ are seed-0 values. a as marked. Arms at $M_\ell = 8$ per layer unless stated.

a	arm	seeds (PPL / LRU_{160} / Belady_{160})	Static	κ
0.02	base, λ=0	54.71/1,120,458/558,511; 54.37/1,117,335/555,661; 54.34/1,086,918/538,267	920,685	1.22
1.0	base, auxiliary every step	53.80/1,196,581/614,138; 55.61/1,192,672/611,664; 53.78/1,200,052/614,943	1,017,359	1.18
0.02	top-4 of 64, λ=0	57.56/393,671/174,350	—	1.33
0.02	base, 8B tokens	41.60/905,877/405,206	—	1.32
1.0	ours λ=0.2, 8B tokens	47.38/115,171/55,621	—	1.31
1.0	ours λ=0.1	58.78/215,395/110,152; 57.90/241,832/120,045	163,562	1.32
1.0	ours λ=0.2	61.19/169,664/83,880; 62.06/156,954/76,741; 60.22/177,880/88,942	130,548	1.30
1.0	ours λ=0.4	62.83/149,439/71,063	115,412	1.29
1.0	ours λ=0.8	63.82/137,388/62,418	100,375	1.37
1.0	switch-only λ=0.5	56.53/502,237/245,318	493,283	1.02
1.0	switch-only λ=1.0	59.11/327,702/152,331; 58.87/321,184/152,077	324,078	1.01
1.0	switch-only λ=1.5	61.30/192,168/89,165	191,486	1.00
1.0	switch-only λ=2.0	62.81/108,624/49,929; 64.09/112,806/52,119	106,451	1.02
1.0	switch-only λ=2.5	66.84/58,433/27,894; 67.08/75,372/33,717	52,004	1.12
1.0	switch-only λ=3.0	69.92/29,050/14,694; 68.87/34,390/16,611	22,190	1.31
1.0	switch-only λ=3.5	70.11/16,084/9,318	9,318	1.73
1.0	BlockFFN analogue, chunk 8, λ=1.0	66.25/122,921/53,545	86,473	1.42
1.0	concentration-only λ=0.2	55.84/402,119/204,336	315,013	1.28
1.0	concentration-only λ=0.5	57.52/257,112/134,888	—	1.25
1.0	retrofit λ=0.2, 10^{-4}, 1.6B	60.06/450,288/216,485; 59.90/444,822/212,823	—	1.38
1.0	retrofit λ=0.2, 4×10^{-4}, 1.6B	53.01/152,762/75,162	—	—
0.3	ours λ=0.1	58.67/170,635/86,298	133,353	1.28
0.3	ours λ=0.2	62.85/80,532/41,063; 61.86/95,723/49,245; 62.31/104,964/53,530	60,877	1.32
0.3	switch-only λ=1.5	63.09/159,465/73,019	149,909	1.06
0.3	switch-only λ=2.0	64.62/70,638/33,293; 64.46/91,223/41,210; 66.48/90,199/41,173	64,103	1.10
0.3	concentration-only λ=0.2	56.56/380,837/189,474	288,003	1.32
0.1	ours λ=0.2	64.16/37,189/19,423; 64.95/30,398/17,180; 66.40/31,591/17,833	23,702	1.57
0.1	residency-only λ=0.2	62.76/63,865/32,714; 63.63/53,630/28,847; 64.05/61,373/32,918	43,885	1.46
0.1	residency-only λ=0.35	66.12/36,926/19,311	23,660	1.56
0.1	switch-only λ=2.0	64.82/79,032/35,935; 65.33/66,475/31,309; 66.08/71,021/32,802	66,405	1.19
0.1	concentration-only λ=0.2	55.97/356,097/175,861; 57.34/344,851/171,207; 57.46/348,089/176,132	270,997	1.31
0.1	concentration + switch λ=0.2	57.28/231,133/119,132	187,862	1.23
0.1	concentration-only λ=1.0	59.19/210,615/109,234	166,763	1.26
0.1	concentration-only λ=2.0	59.89/197,549/104,237; 60.37/178,571/92,833; 61.75/171,838/89,926	159,846	1.24
0.1	concentration + switch λ=2.0	61.33/149,633/76,633	—	—
0.1	concentration + switch λ=5.0	63.12/142,887/73,599	—	—
0.01	ours λ=0.1	61.75/62,607/32,042	46,452	1.35
0.01	ours λ=0.125	62.59/39,315/20,991	—	1.43
0.01	ours λ=0.15	64.64/27,141/14,868; 64.58/29,979/16,550; 66.11/25,511/14,251	17,563	1.55
0.01	ours λ=0.2	66.43/14,152/8,755; 67.32/15,530/9,461; 66.79/13,007/8,271	7,467	1.90
0.01	residency-only λ=0.15	63.18/43,285/22,719	30,425	1.42
0.01	residency-only λ=0.25	65.94/16,991/10,141; 65.67/23,460/13,307; 66.53/22,176/12,832	9,771	1.74
0.01	memoryless M_ℓ=16, λ=0.15	59.62/379,050/129,042	238,353	1.59
0.01	stateful λ=2.0, M=160	60.94/50,473/19,530	—	—
0.01	stateful λ=4.0, M=160	61.22/11,243/7,873; 63.51/11,289/7,957; 63.42/10,341/7,356	5,928	1.90
0.01	stateful λ=4.0, M=80	64.12/27,671/13,804	17,946	1.54
0.01	switch-only λ=1.5	65.36/63,260/29,693; 64.47/79,200/36,946; 65.62/68,675/32,917	50,329	1.26
0.01	switch-only λ=2.0	65.49/76,230/34,778	61,498	1.24
0.01	switch-only λ=3.0	67.97/20,417/10,588; 69.47/19,347/10,344; 68.57/14,422/8,683	12,581	1.62
0.01	switch-only λ=4.0	69.15/5,490/4,729	1,167	4.70
0.01	squared-ℓ_2 $\lambda \in \{2, 4, 16, 64\}$ (one run each)	56.42/616,660/291,358; 57.19/513,424/240,827; 58.76/305,996/142,853; 61.75/122,597/55,461	560,596; 486,440; 320,064; 136,629	1.10; 1.06; 0.96; 0.90
0.01	retrofit λ=0.2, 10^{-4}, 1.6B	61.60/445,646/214,181	327,470	1.36
0.01	retrofit λ=0.5, 10^{-4}, 1.6B	89.34/90,622/47,718	70,337	1.29

Table 6: κ decomposition (Eq. 3) at 414M/64 experts, capacity 160, seed 0. $(1-h_c)/(1-q)$ drops the boundary term; g_4 is the fraction of cold accesses whose previous access is at most 4 tokens back; Sw_1 is split by whether the entering block is hot or cold. Switch-trained arms sit at $q \approx h_c$, marginal-trained arms at $q > h_c$.

a	arm	PPL	Static	κ	$\frac{1-h_c}{1-q}$	h_c	q	g_4	Sw_1 hot / cold
0.02	base	54.71	920,685	1.22	1.22	.29	.42	.43	634,749 / 735,519
1.0	ours λ=0.2	61.19	130,548	1.30	1.28	.30	.46	.22	434,910 / 114,939
1.0	ours λ=0.8	63.82	100,375	1.37	1.35	.36	.52	.23	331,654 / 88,422
1.0	switch-only λ=1.0	59.11	324,078	1.01	1.01	.49	.49	.48	496,094 / 217,122
1.0	switch-only λ=2.0	62.81	106,451	1.02	1.00	.52	.52	.40	382,641 / 76,552
1.0	switch-only λ=3.0	69.92	22,190	1.31	1.17	.46	.54	.26	313,374 / 18,004
1.0	concentration-only λ=0.2	55.84	315,013	1.28	1.27	.26	.42	.27	492,177 / 272,586
1.0	BlockFFN analogue	66.25	86,473	1.42	1.39	.35	.54	.23	464,867 / 77,422
0.1	ours λ=0.2	64.16	23,702	1.57	1.47	.28	.51	.10	365,503 / 22,275
0.1	residency-only λ=0.2	62.76	43,885	1.46	1.39	.26	.47	.10	416,550 / 41,492
0.1	switch-only λ=2.0	64.82	66,405	1.19	1.15	.47	.54	.31	340,347 / 52,454
0.1	concentration-only λ=2.0	59.89	159,846	1.24	1.22	.26	.40	.18	302,033 / 147,334
0.01	ours λ=0.2	66.43	7,467	1.90	1.53	.27	.52	.04	293,611 / 7,260
0.01	ours λ=0.1	61.75	46,452	1.35	1.29	.30	.46	.12	388,591 / 43,508
0.01	switch-only λ=1.5	65.36	50,329	1.26	1.20	.43	.52	.25	295,984 / 41,468
0.01	switch-only λ=2.0	65.49	61,498	1.24	1.20	.44	.53	.27	300,172 / 50,845
0.01	retrofit λ=0.2, 1.6B	61.60	327,470	1.36	1.36	.27	.46	.30	628,570 / 276,611

C.3 Scaling ladder and robustness axes

Table 7: Scaling ladder at $a = 1.0$, identical 2B-token protocol, misses at the 25% tier. Multiples are base over born (LRU / Belady) on seed 0, with the seed-mean LRU multiple in parentheses where seeds exist (seeds base/born in the "n" column); "retrofit" is the multiple of the 200M-token 10^{-4} retrofit at λ=0.2 and "gap" the Belady ratio between the budget-matched 10^{-4} retrofit and the born arm, a protocol number that closes at 520M and 414M when the retrofit uses the pretraining rate (Section 5.3); the 1.3B gap is against the 200M retrofit and no budget-matched retrofit exists at 6.8B. Bottom, the token and domain axes (single runs). S1 sits at the routing sparsity floor. Duty-matched bases raise the 520M and 414M multiples to 3.55× and 7.11× (Appendix A.3).

model	geometry	n	λ=0.05 mult.	λ=0.2 mult.	rel. ΔPPL	retrofit mult.	gap
S1 62M	8L × 16E, top-4	3/3	1.06× / 1.12×	1.15× / 1.33× (1.17)	+12.6%	—	—
S2 225M	10L × 32E, top-4	3/2	1.55× / 1.90×	2.99× / 4.21× (3.01)	+8.9%	—	—
S3 520M	12L × 32E, top-4	3/3	1.75× / 2.15×	3.22× / 4.51× (3.13)	+12.8%	1.40× / 1.63×	1.97×
G 414M	10L × 64E, top-8	3/3	—	6.60× / 6.66× (6.6)	+11.8%	—	2.58×
S4 1.3B	16L × 32E, top-4	1/1	—	3.11× / 4.37×	+9.8%	1.50× / 1.72×	2.54× (200M)
S5 3.6B	18L × 64E, top-8	2/2	—	15.0× / 13.5× (16.2)	+9.3%	2.80× / 2.70×	3.08×
S6 6.8B	16L × 64E, top-8	2/2	—	13.7× / 11.9× (13.8)	+12.9%	2.46× / 2.40×	—
S3 520M, 8B tokens	base PPL 35.71	1/1	—	3.23× / 4.17×	+9.0%	1.60× / 1.84× (800M)	—
G 414M, 8B tokens	base PPL 41.60	1/1	—	7.87× / 7.29×	+13.9%	—	—
S3 520M, C4 traces	out of domain	1/1	1.77× / 2.21×	3.32× / 4.62×	—	1.49× / 1.78×	—

Per-seed multiples at 3.6B are 15.0× / 13.5× and 17.5× / 15.4× (relative ΔPPL +9.3% and +7.8%), at 6.8B 13.7× / 11.9× and 14.0× / 11.7× (+12.9%, +6.8%); at 414M the paired LRU multiples are 6.60, 7.12 and 6.11×. The budget-matched 10^{-4} retrofit at 3.6B reaches 4.75× / 4.39× at +6.3%. Every retrofit above 520M used the 10^{-4} protocol; the 4×10^{-4} retrofit that closes the gap at 520M and 414M was not run at 1.3B, 3.6B or 6.8B.

C.4 Downstream quality

Table 8: Zero-shot battery, lm-eval 0.4.12, `acc` in percent over ARC-Easy, HellaSwag, LAMBADA, PIQA, SciQ and WinoGrande; one run per arm except the two three-seed rows, whose per-seed means are listed. LRU multiples in parentheses are those of Table 7. The two $a = 0.01$ rows are the matched cell of Section 5.1; per-task means ours/switch-only: LAMBADA 14.8/15.3, HellaSwag 27.2/27.1, PIQA 59.1/58.5, ARC-Easy 44.3/44.0, WinoGrande 51.6/51.9, SciQ 62.2/62.9. Right, OLMoE-1B-7B (`acc_norm` where defined) with only the router changed.

arm	ΔPPL	mean acc
520M base	—	46.14
520M born λ=0.05 (1.75×)	+2.2	45.86
520M retrofit λ=0.2, 10^{-4} (1.40×)	+4.7	45.54
520M born λ=0.2 (3.22×)	+5.8	45.29
520M top-2 of 32 (2.93×)	+3.2	44.76
414M/64E base	—	43.43
414M/64E top-4 of 64 (2.85×)	+2.9	43.29
414M/64E born λ=0.2 (6.60×)	+6.5	43.40
414M/64E, a=0.01, ours λ=0.15 (43.03/43.20/43.31)	—	43.18
414M/64E, a=0.01, switch-only λ=1.5 (43.11/43.53/43.14)	—	43.26
1.3B base	—	47.99
1.3B born λ=0.2 (3.11×)	+3.7	47.41
6.8B base	—	51.52
6.8B born λ=0.2 (13.7×)	+3.9	50.51

task	base	λ=0 floor	adapted
ARC-Challenge	48.72	47.95	47.61
ARC-Easy	76.14	75.59	75.76
BoolQ	74.46	74.89	75.50
HellaSwag	76.95	76.84	76.58
PIQA	80.63	80.41	80.69
WinoGrande	68.35	68.03	68.59
mean	70.87	70.62	70.79

C.5 Released-model adaptation

Over the 24×1024-token base traces mean LRU/Belady is 1.75 (1.61, 1.68, 1.96 at 6, 12.5 and 25%) and the best static pin trails Belady everywhere (static/Belady 1.29, 1.52, 1.86), so the headroom is temporal rather than pinnable skew; the evolved cache policy closes 0.61, 0.36 and 0.21 of the gap at the three tiers. Tables 9 to 11 give the head-to-head under one guard, the router × cache factorial, the loss-budget sweep, the stateful comparison and the H100 replay.

Table 9: Head-to-head under one guard (OLMoE router fine-tuning, 2M tokens; gain is the reduction in cache misses; guard is worst-seed ΔPPL ≤ 0.3 against the fine-tuning floor). The switch-only control is tuned over a $20\times$ λ range reaching past its own quality cliff. Differences among the valid arms are not significant (Welch $p \geq 0.34$; miss-only over switch-only λ=1.5 is +0.011, 95% interval $[-0.050, 0.073]$).

arm	n	gain (mean±std)	ΔPPL mean/max	guard viol.
switch-only λ=0.1–0.5	10 ea	+0.019 – +0.038	$\leq$ +0.09	0
switch-only λ=1.0	10	$+0.074 \pm 0.049$	+0.107/+0.199	0
switch-only λ=1.25	10	$+0.104 \pm 0.051$	+0.149/+0.190	0
switch-only λ=1.5 (best valid)	10	$+0.105 \pm 0.079$	+0.209/+0.286	0
switch-only λ=2.0	10	$+0.148 \pm 0.071$	+0.352/+0.428	10/10 (invalid)
ours miss-only λ=0.15	20	$+0.116 \pm 0.064$	+0.179/+0.238	0
ours miss-only, M_ℓ=4, λ=0.15	10	$+0.123 \pm 0.037$	+0.142/+0.189	0
ours full λ=0.18 (searched recipe)	9	$+0.111 \pm 0.033$	+0.289/+0.317	4/9
stateful W=64, λ=0.25	20	$+0.104 \pm 0.048$	+0.243/—	2/20

The wall-clock figures of Section 5.4 come from a longer replay over 2560 tokens in which the adapted router moves 0.63 against 0.80 GB per token at the 25% tier; a shorter transfer-only replay

Table 10: Left, router × cache factorial on OLMoE (prefetch tax 0.25; gains are fractional cost reduction against LRU on the un-adapted workload; three decode seeds per arm, bootstrap 95% intervals over run × chunk cells; synergy is joint minus cache minus router; the 6%-tier synergy is +12.8/+13.0/+13.2% at prefetch taxes 0.1/0.25/0.5). Right, loss budget M_ℓ against deployment tier (λ=0.15 miss-only; n=5 for $M_\ell \in \{4, 16\}$, the M_ℓ=8 row is the n=20 fleet; mean ± std misses): the tight budget dominates at every tier including M_ℓ=16's own (+14.1% at capacity 256, Welch p=0.016). The stateful window sweep at λ=0.15 (n=2) gives gains of −0.014, +0.002, +0.026, +0.026 at $W = 1, 8, 64, 512$; stateful arms at $\lambda \in \{0.35, 0.5, 0.7\}$ reach +0.149, +0.281, +0.409 with maximum ΔPPL +0.53, +1.23, +2.92, all outside the guard.

capacity	cache	router	joint	synergy (95% CI)
64 (6%)	+24.4%	+0.0%	+37.4%	+13.0% [+5.9, +21.2]
128 (12%)	+15.1%	+18.8%	+32.7%	−1.1% [−3.6, +0.9]
256 (25%)	+11.9%	+26.7%	+37.4%	−1.2% [−3.7, +1.3]

M_ℓ	miss@64	miss@128	miss@256	ΔPPL
4	264,143 ± 10,749	212,046 ± 10,824	134,340 ± 8,253	+0.160
8	272,290 ± 16,100	219,999 ± 16,129	139,037 ± 13,791	+0.179
16	284,353 ± 16,675	232,099 ± 16,698	156,332 ± 12,997	+0.187

Table 11: H100 replay of OLMoE-1B-7B traces (pinned host to GPU, real 12.6 MB blocks), base against the adapted router, LRU and the evolved policy. Prefetch copies share the default CUDA stream, so the evolved policy's rows at capacity 64 are penalized by stream serialization and its stall numbers understate a dual-stream implementation; the LRU rows are unaffected. Each row is one replay.

traces / cap	policy	GB/token	stall ms	traces / cap	policy	GB/token	stall ms
base 64	LRU	1.613	65.4	adapted 64	LRU	1.613	65.2
base 64	evolved	1.917	71.0	adapted 64	evolved	1.680	61.3
base 128	LRU	1.181	48.0	adapted 128	LRU	0.970	39.2
base 128	evolved	1.012	46.2	adapted 128	evolved	0.808	36.3
base 256	LRU	0.803	32.5	adapted 256	LRU	0.454	18.3
base 256	evolved	0.730	39.5	adapted 256	evolved	0.395	21.4

over the first 1536 tokens gave −43% on an anomalous first chunk and is not quoted. The differentiable proxy stays rank-faithful to the simulator across 79 adapted workloads (Spearman 0.68–0.81), with one ceiling-effect blind spot (LRU at 6%).

D MEASUREMENT AND REPRODUCIBILITY

Blocks and traces. A block is one expert's weight tensor for one layer. A trace is the per-token sequence of accessed blocks (top-k per layer × layers accesses per token: 128 for OLMoE-1B-7B and the 6.8B replica, 80 for the 64-expert 414M geometry, 48 for 520M). Pretrained arms are traced on 24 chunks of 1024 WikiText-103 test tokens under teacher forcing (1,179,648 accesses at 520M, 1,966,080 at 414M/64E), validated against decode traces on 12 chunks (rank correlation 1.00); every adapted released-model workload is evaluated on its own freshly recorded decode traces. C4 traces use the same chunking on C4 validation text. GB per token is misses per token × 12.6 MB.

Simulator. `src/hermit/sim/cache.py`. LRU and Belady (offline furthest-in-future replacement (Belady, 1966)) run with mandatory insertion from a cold cache at the start of each chunk; Static_M pins each chunk's M most frequent blocks and serves the rest by bypass with no fill. Capacities are counted in blocks and quoted as a fraction of all blocks. The released-model policy score is cost = demand misses + 0.25 × prefetches (0.1 and 0.5 as robustness profiles); LRU, LFU and Belady issue no prefetches, so their cost is the miss count. The "gain" of a released-model arm is the fractional reduction of the evolved policy's summed cost over capacities $\{64, 128, 256\}$ on the arm's own traces against the same quantity on the un-adapted model's traces, and "fraction of the gap closed" for a policy P is $(\text{cost}_{\text{LRU}} - \text{cost}_P)/(\text{cost}_{\text{LRU}} - \text{cost}_{\text{Belady}})$ over 24 chunks.

Quality. WikiText-103 test perplexity at sequence length 1024; downstream scores from lm-eval 0.4.12 (`acc_norm` where defined, else `acc`), batch 16, one protocol for every arm. ΔPPL for released-model arms is against the λ=0 fine-tuning floor and guard statements are always worst-seed.

Hardware replay. `scripts/tier_transfer_bench.py` holds the M resident blocks on an H100 and copies each missed block from pinned host memory ($\approx$ 25 GB/s effective, consistent with pinned PCIe), either before the token's forward pass (blocking) or on a concurrent stream overlapped with it (an upper bound). Throughput changes are against the base model's traces through the same harness. Measured compute for the 6.8B replica is 19.9 ms per token.

Aggregation. A multiple is base misses over arm misses at one capacity and policy, on seed means where seeds exist; a quality-matched ratio is arm over control at the closest mean perplexities; percentages of PPL are relative to each arm's own base. Welch t-tests are two-sided over seeds; bootstrap intervals for the factorial are over run $\times$ chunk cells.

Artifacts. Fidelity records for every arm are `data/wsb_fidelity_<arm>.json` (perplexity and LRU, LFU and Belady misses per capacity); `data/arm_inventory.csv` lists every arm of Appendix C and the omitted negatives with these values; `data/static_kappa.json`, `data/kappa_vs_cap.json` and `data/kappa_decomp.json` hold the κ analyses (`scripts/static_kappa.py`, `scripts/kappa_vs_cap.py`, `scripts/kappa_decomp.py`); training recipes and submitters are under `experiments/04_pretrain_codesign` (pretraining) and `experiments/03_coadapt_real` (released-model adaptation); figures regenerate from `paper/figs/make_fig_*.py`; the pre-registration and outcome records are in `docs/OBJECTIVE_V2.md`.