D-model and D-Module: Identifying and Eliminating Model-, Task-, and Request-Level Transformer Redundancy

Anonymous authors
Paper under double-blind review

Abstract

Which computation in a trained Transformer can be removed while improving quality and reducing inference cost? We study this question under a fixed local compensation budget. Dynamic H couples D-model, which predicts the gains of a trained replacement, with D-Module, a causal rank-96 bridge trained through the frozen model's suffix. The same mechanism supports model-, task-, and request-level reuse; only request-level recognition incurs per-request cost. Twenty of 24 fixed crossings across three Qwen backbones improve observed mean task CE, accuracy, and model-path time. A matched control improves quality without removing layers but adds latency, separating adaptation's quality benefit from removal's execution benefit. The same training and selection procedure produces joint-positive actions on Qwen3, SmolLM2, and Pythia. Automatic task selection and request routing also retain joint gains. Shared-checkpoint model reuse improves mixture CE and time at unchanged accuracy, although task effects oppose each other. These findings establish useful compensated removals and procedural transfer across model families. They also distinguish bridge recoverability from selector quality: the learned selector finds usable actions, but composite-utility superiority over geometric selection is not established.

Keywords: Transformer pruning, model compression, layer replacement, adaptive inference, task-aware model reuse, causal bridge adaptation.

1 Introduction

Trained Transformer language models contain computation that can be removed without proportionate loss of capability [1]. Realizing that redundancy matters because the same deep backbone is repeatedly executed across tasks and requests: a cheaper path that preserves or improves final quality can reduce inference work. Yet redundancy is not simply a property of layer count or hidden-state similarity. Removing an interval changes the state entering the remaining model; recovery can restore task performance but consume the saved time. The scientifically relevant object is therefore what can be removed under a limited compensation rule, with what final quality and runtime, and over which workload that benefit can be reused.

Layer pruning and replacement already identify less important computation and recover performance [1–3]. Streamline also trains replacement modules with a language-model objective while freezing the original model [2]. Surrogates estimate pruning responses [4], search learns task-specific routes [5], and input-adaptive routers select conditional paths [6, 7]; calibration can change observed redundancy [8]. The unresolved issue we study is how identification relates to the benefit of an installed, budget-constrained replacement. Original-layer importance, a powerful teacher's response, and a lightweight bridge's realized benefit are different targets. Quality recovery alone is insufficient if compensation consumes the saved execution time. Our predictor therefore estimates the fitted student's final-task and runtime gains, and its reuse policy is evaluated on that same action.

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

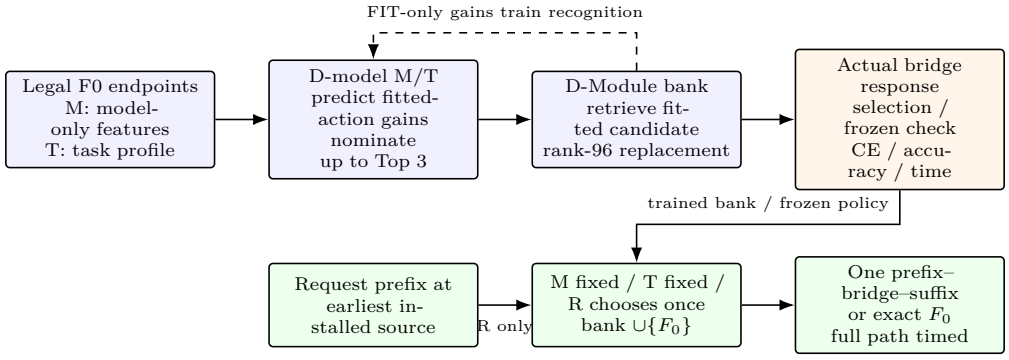

Figure 1: Dynamic H development and execution. The bridge bank is post-trained and labelled before D-model fitting; only FIT gains supervise recognition. Predictions nominate bank candidates for selection and a frozen-policy check. M reuses one source-distribution action, T one task action, and R chooses once from a frozen bank using a legal prefix. Only R pays online recognition. The enumeration spans the model depth; automatic gain tests cover nine measured sites.

We formulate realizable redundancy under limited compensation. A fixed local bridge and training recipe replace a selected interval of the frozen backbone. D-Module is trained through the original suffix to preserve the state components needed for the task output, rather than relying on local state similarity alone. D-model predicts this executable replacement's quality and time gains. Final-output observations supervise adaptation offline; deployment omits the counterfactual observer.

The fitted action can be reused across a declared task family (model level), across one task's requests (task level), or selected conditionally from a trained bank (request level). These three horizons answer different redundancy questions. A single checkpoint must retain useful gains across tasks for model reuse; a task-fixed path must work across requests; a request policy must recover enough conditional benefit to pay for recognition. We call the action together with its policy and workload a removal unit.

This leads to three linked questions. Realizability: can limited local compensation produce faster, better final outputs rather than quality recovery alone? Identifiability: can execution-before features find those opportunities better than simple structural or teacher-based controls? Reusability: how far do realized gains survive changing task, backbone, or decision granularity? We study these questions with the same intervention semantics and component-wise measurements, without treating success at one horizon as success at all three.

Our contribution has three parts. Mechanism: D-model learns the realized quality–time response of bounded D-Module compensation, with information and execution costs defined separately for M/T/R. Realization: rank-96 bridges yield joint gains at 20 of 24 Qwen crossings; matched adaptation controls distinguish quality recovery from shorter execution. Reuse: the same bridge-training and selection procedure produces joint-positive actions across Qwen, SmolLM2, and Pythia, while automatic task and paid request decisions provide usable anchor instances. The comparative identification question remains open: geometry has higher mean composite utility, and native importance selects the same task action. Shared-checkpoint testing further exposes task cancellation. Together, these results explain when bounded compensation realizes removable computation and where identifying or reusing it remains difficult.

2 Problem: Realizable Redundancy under Limited Compensation

We seek computation in a trained model whose removal, with bounded local compensation, improves the quality–cost outcome on a declared workload. The compensation budget fixes the bridge structure, trainable parameters per site, optimization recipe, and number of

updates; it does not permit retraining the backbone. Redundancy is operational: it is relative to that intervention class, training rule, workload, and execution regime. It is not a claim that an untouched interval is intrinsically useless or that unlimited retraining could reconstruct the entire model.

Let F_0^m be a frozen decoder with D_m Transformer blocks. We index layer boundaries from zero: h_s enters block s, and $F_{u:v}^m$ executes blocks $u, \ldots, v-1$. Task τ specifies a request distribution, labels, readout q_τ, and loss L_τ. A legal pair $e = (s, d)$ satisfies $0 \le s < d < D_m$ and the backbone's state, position, and cache constraints. An action a fixes the pair and fitted bridge B_a. It replaces interval $[s, d)$:

$$F_a^m(x) = F_{d:D_m}^m(B_a(F_{0:s}^m(x))), \qquad e = (s, d), \tag{1}$$

where the prefix includes input embedding and the suffix ends at the final backbone state. The task readout follows as $q_\tau(F_a^m(x))$. A removal unit adds a policy, workload, and timing regime to the action. It executes one prefix–bridge–suffix path or exact F_0^m fallback, without evaluating counterfactual paths. For a request policy π, $F_\pi^m(x)$ denotes its chosen path.

For a request x with label y, define paired final-output and model-path gains by

$$G^{\mathrm{CE}}(a; x) = L_\tau(q_\tau(F_0^m(x)), y) - L_\tau(q_\tau(F_a^m(x)), y), \tag{2}$$

$$G^{\mathrm{Acc}}(a; x) = \mathbf{1}[\hat{y}_a(x) = y] - \mathbf{1}[\hat{y}_0(x) = y], \tag{3}$$

$$G_\omega^{\mathrm{run}}(a; x) = c_\omega(F_0^m, x) - c_\omega(F_a^m, x). \tag{4}$$

Here ω names a timing regime, distinct from destination d. The primary c_ω times pretokenized input to final backbone state, including the bridge, suffix, and online recognition or dispatch. Offline profiling, training, counterfactuals, and task-score observation are outside both timers. This is a backbone-path comparison, not complete task or service latency. For multiple-choice tasks, q_τ computes each option's mean continuation-token log likelihood; CE is cross-entropy over the softmax of these option scores, and accuracy uses their argmax (Appendix A.1). It is not token-level language-model CE. Quality and runtime are measured separately; skipped-layer count alone determines neither.

Objective and empirical gate. Write $\boldsymbol{\mu}(\pi) = (\mathbb{E}G^{\mathrm{CE}}, \mathbb{E}G^{\mathrm{Acc}}, \mathbb{E}G_\omega^{\mathrm{run}})$. Component-wise nonnegative gains with positive composite utility constitute effective improvement: unchanged accuracy is compatible with lower CE and faster execution. A positive utility with a negative component does not imply non-degradation. The stricter research target is

$$\mathbb{E}G^{\mathrm{CE}}(\pi; X) > 0, \quad \mathbb{E}G^{\mathrm{Acc}}(\pi; X) > 0, \quad \mathbb{E}G_\omega^{\mathrm{run}}(\pi; X) > 0. \tag{5}$$

The implementation estimates these expectations by paired sample means and admits a frozen policy only when all three observed means are strictly positive. This empirical gate is not a statistical certificate of Equation (5); descriptive intervals accompany it. Within the nominated pool, selection maximizes the fixed utility in Equation (8), not the unknown global population objective. M averages over declared source-task weights; T/R average over one task's requests. Mixtures retain per-task effects. Failed qualification returns F_0, whereas useful aggregate benefit without strict qualification is reported separately.

Model-, task-, and request-level redundancy are three reuse horizons, with distinct information, granularity, and cost. At the model level, task-neutral architecture and unlabeled probes provide candidate inputs; a predictor learns source-task-mixture gains and nominates a model-fixed crossing. Its bridge is fitted on that declared distribution, without target-specific inputs to the predictor. At the task level, labels and gradients observed offline select a task-fixed crossing. Both discovery costs are paid during adaptation, not per deployed request. At the request level, metadata and one prefix state available before the earliest candidate source choose among installed actions or F_0; that inference cost is charged on every request, including fallback. R may not inspect candidate suffixes or final answers before choosing. The research question is how to identify bounded-compensation opportunities, realize their final-task and runtime benefits, and reuse them across these horizons under the same rules. A candidate may benefit an identifiable request subset despite failing task-wide admission. R therefore evaluates the routed policy's means, not each candidate's unconditional means. Success at one horizon does not establish another.

3 Dynamic H: Identifying, Realizing, and Reusing Removals

Dynamic H couples candidate identification to a fixed compensation procedure. D-Module fits replacements; their observed CE, accuracy, and runtime gains supervise D-model. Predicted gains nominate candidates, measured gains select an action, and a later check tests the frozen policy. This separates learning what is recoverable from deciding where to deploy it.

3.1 D-Module: bounded compensation through the original suffix

For each source, an offline teacher has short and long transport branches. Source residual adaptation and causal attention form each direct branch; the long branch also attends to a learned short-destination state. Four stages train state/dynamics reconstruction, task response, hierarchical reconstruction, and hierarchical task response. A causal rank-96 student distils the selected destination and is then trained through the unchanged suffix. Only bridge parameters receive updates. Each crossing has independent student weights. M uses its declared source-task mixture; T uses task-development data; R reuses T bridges. Appendix A.1 defines the teacher and all losses.

With $z_{s,t} = \mathrm{LN}(h_{s,t})$ and masked causal prefix mean $\bar{z}_{s,\le t}$, the bridge is

$$B_{s,d}(h_s)_t = h_{s,t} + m_t W_{\mathrm{up}}^{s,d}\, \phi\Big(W_{\mathrm{down}}^{s,d}[z_{s,t}; \bar{z}_{s,\le t}]\Big), \tag{6}$$

where ϕ is SiLU, m_t masks padding, and bias-free projections map $2H \to 96 \to H$, with zero-initialized up-projection. Student training uses 256 hidden-distillation and 512 final-task steps with loss $\mathrm{CE} + 0.2\,\mathrm{MSE}_{\text{centered score}} + 0.05\,\mathrm{error}_{\text{hidden}}$; stop-role CE chooses the checkpoint. The teacher stages have 768/768/384/384 updates. M balances tasks within these total budgets. The fitter requires $3 \le s < d_{\mathrm{short}} < d_{\mathrm{long}} < D_m$; enumeration alone does not establish training feasibility.

Execution replaces blocks $s, \dots, d-1$ by one bridge and continues through original blocks $d, \dots, D_m - 1$. The bridge runs directly; its arithmetic is included in the model-path timer.

Why destination similarity is insufficient. Let $\ell_x(u)$ be the final-task loss after the unchanged suffix, and let $\delta_a = B_a(h_s) - h_d$ be the destination-state error. Locally,

$$G^{\mathrm{CE}}(a; x) = -\langle \nabla \ell_x(h_d), \delta_a \rangle - r_a, \qquad |r_a| \le \tfrac{\beta_x}{2}\|\delta_a\|^2, \tag{7}$$

when the loss Hessian is bounded by β_x along that segment. Hence error magnitude alone does not determine gain; its alignment with downstream task sensitivity also matters. This standard local expansion motivates task-effect features and suffix-based training, not a guarantee of positive CE, accuracy, or time (Appendix A.2).

3.2 D-model: predicting the gains of that compensation rule

D-model enumerates architecture-legal endpoint pairs offline. Its coordinate scan is $O(D_m^2)$, not a request-time loop. Model features describe endpoint fractions, span/suffix geometry, width, head/KV ratios, and F0 states from a fixed task-neutral probe. Task features add F0-only displacement, gradient–effect, margin, confidence, and output-form summaries. These features precede the candidate intervention; the labels are paired post-trained student CE, accuracy, and time outcomes on FIT identities.

Separate heads estimate $\widehat{\boldsymbol{g}}^M(e)$, $\widehat{\boldsymbol{g}}^T(e)$, and $\widehat{\boldsymbol{g}}^R(a, x)$. M labels aggregate the declared source-task mixture without supplying task-specific predictor inputs; T retains task identity; R retains each candidate's outcome on the same request. The target is thus benefit recoverable by this bridge recipe, not an unconstrained prediction of redundancy. For a selection role $\mathcal{S}$,

$$U_{\mathcal{S}}(a) = \frac{\bar{g}_{\mathcal{S}}^{\mathrm{CE}}(a)}{0.1} + \frac{\bar{g}_{\mathcal{S}}^{\mathrm{Acc,pp}}(a)}{5} + \frac{\bar{g}_{\mathcal{S}}^{\mathrm{run,ms}}(a)}{1}, \qquad a_{\mathcal{S}}^* = \arg\max_{a \in \mathcal{B}_{\mathcal{S}}^+} U_{\mathcal{S}}(a). \tag{8}$$

$\mathcal{B}_{\mathcal{S}}^+$ contains candidates with all three measured means positive; an empty set returns F_0. The scales encode a fixed trade-off, not a universal quality unit. For M/T, predicted $\widehat{U}$

nominates at most three predicted triple-positive sites; measured U selects a frozen Top-1, and another role evaluates it. Best means utility-best within the qualified candidate pool, not the longest span, largest CE gain, or an unmeasured global optimum.

The measured-pool M/T implementation uses standardized ridge heads ($\alpha = 10$). Each source is predicted using other-source FIT labels; all predictions compete for Top 3. M uses 18 task-neutral features versus four geometry features; T uses 16 structural/task-effect features and interactions versus six geometry/teacher-coordinate features. The richer inputs let nomination condition on model states and task effects, not only endpoint geometry; candidate-source outcomes do not train that source's head. Full-depth enumeration is an interface capability, while completed gain prediction and qualification cover nine anchor sites.

For R, an additive term $b(a) + c(x)$ cannot change the relative ranking of two actions, although it may change threshold eligibility. We use candidate–request interactions: $\widehat{\boldsymbol{g}}^R(a, x) = \boldsymbol{b}_a + (W_0 + \sum_{j=1}^{4} \gamma_j(a) W_j)\psi(h_{s_{\min}})$, where γ encodes geometry and ψ summarizes source mean/last states, norms, and length in 68 dimensions, invariant to option ordering. Supplied option texts are allowed; correct labels, gradients, later states, and candidate outputs are not. R maximizes scalar $\widehat{U}(a, x)$, returning F_0 if the maximum is at most $3 \cdot 10^{-12}$ (numerical tolerance). Component-wise positivity qualifies aggregate policy means, not each request prediction. Static terms compile offline; one prefix summary and at most three predictions run online.

3.3 Reuse horizons and qualification

M installs one source-distribution-trained checkpoint across declared tasks. T installs a task-fixed action. R chooses among frozen task candidates or F_0 at the earliest source $s_{\min}$; a later-source action continues unchanged to its own source. M/T discovery and training are offline. Every R request, including fallback, pays feature extraction, head inference, and dispatch.

Selection freezes the action or policy before internal checking. Installation requires the original strict three-mean gate; the check cannot reselect a winner. R may use an unconditionally weak candidate if the resulting policy qualifies. Model mixtures retain per-task effects. External evaluation cannot alter the frozen mechanism. The three scope-specific implementations are not a demonstrated joint horizon controller.

Offline development and online execution. First fit and label a bridge library, then train source-crossfitted gain heads on FIT labels. Selection measures up to three nominees and freezes the qualified action or request policy; checking cannot reselect. Deployment executes its frozen path or bank without the observer. These tests concern a preconstructed library, not reduced full-depth acquisition cost.

4 Experimental Design

The experiments test the removal unit's realization, identification, and reuse. Local same-site controls test compensation; same-library controls test selection; within- and across-family experiments test procedural transferability. These complementary evidence levels retain their own checkpoints, panels, and timing windows.

Common objects and measurements. The anchor is Qwen3-0.6B/HellaSwag. Other backbones are Qwen1.5-1.8B, Qwen2.5-1.5B, SmolLM2-1.7B, and Pythia-1.4B; BoolQ is M's two-choice source task. Fit, checkpoint-stop, action-selection, and internal-check identities are disjoint within each run. These are development evaluations: their identities have been inspected in earlier work, not reserved as new external tests. Automatic checks use seed 1 and same-device RTX 5060 Ti pairing. We report paired task CE, accuracy points, saved milliseconds, empirical admission, U, and descriptive 95% intervals. Utility compares alternatives on one workload, not different backbones.

BF16/eager, pretokenized batch-1 no-cache timing includes model-input preparation, prefix, bridge, suffix, and R's controller. Batch 1 means one question, evaluated as four option sequences for HellaSwag or two for BoolQ, not one token sequence. Final-score observation and teacher/F0 counterfactuals are outside both timers. F0–action–F0 brackets and alternating order limit drift. M uses fixed task weights and stratified variance. Appendix A specifies data roles and timing.

4.1 Q1: Is compensated removal realizable?

Eight fixed-source spans per Qwen3/Qwen1.5/Qwen2.5 backbone test independently trained rank-96 bridges at equal per-site budgets. Each site has two 512-request roles; only identical site checkpoints are pooled. This is a breadth check, not selection.

At frozen $13 \to 14$, a no-skip residual control retains original blocks. It matches the bridge architecture, source inputs, teacher, seed, data, losses, optimizer, 256+512 student steps, and 296,960 trainable parameters. Separately fitted weights make this a matched-procedure comparison, not deletion at identical weights. Selection 128 precedes internal 512, and both executors pay their actual work in one timing window.

A second fixed-site control ports Streamline's released MLP/MSE alternative [2] to Qwen3: biased $h \to 4h \to h$ ReLU, original-state MSE, and its AdamW/cosine recipe. Both students receive 768 updates and the same data roles. Architecture, teacher access, objectives, parameter count, and total development cost differ. The comparison tests the complete replacement recipes at this site, not the necessity of any one ingredient. Stop MSE selects the port checkpoint; selection freezes port/F_0 before internal testing. This is a released-alternative segment comparison, not Streamline's flagship inherited-Transformer or full algorithm.

4.2 Q2: Can the executable opportunities be identified?

Qwen3 has 378 endpoint feature pairs but nine measured sites (sources 10/12/13, spans 1/3/5). T labels use 32 FIT requests/site; true M labels use 32/task/site from the same mixture-trained checkpoint. M's task-neutral probe uses 16 sequences of 64 integer tokens. Each held-source ridge head uses six other-source sites, with fixed $\alpha = 10$ and utility scales; all predictions compete for at most three candidates.

On the same T bridge library, compare executed-student gain labels, exact-teacher quality labels with the same student-time target, geometry, and native normalized interval-displacement importance. Equal Top-3 budgets and 128 selection/512 paired internal requests isolate nomination while keeping execution fixed. M compares neutral and geometry heads under the same permissions and library. These tests measure selection quality in a common pool, not which framework has broader scope or depth coverage. Retrospective admitted-best recovery/regret describes only that pool; three-site Top-3 pools are trivial. A separate 39-site/five-backbone T check excludes all Qwen3 labels.

FIT/STOP diagnosis compares one frozen M student with its exact teacher on 32 FIT and 64 STOP requests/task. R diagnostics use 32 FIT requests and contrast full routing with the identical forced action. Neither diagnostic changes the head, checkpoint, threshold, or frozen choice.

4.3 Q3: Across what horizons can gains be reused?

Across-family task adaptation uses $s = \lfloor D/2 \rfloor - 2$, offsets $1, 3, 5$ in the middle third, on Qwen3/SmolLM2/Pythia. Independent site bridges follow the same teacher/student rule; 128 selection freezes Top-1/F_0 before 512 internal. This development-revised geometry tests procedure reuse with model-specific weights, not automatic full-depth selection or direct weight transfer.

Model reuse instead trains one checkpoint per site over a 50:50 HellaSwag/BoolQ distribution, balancing tasks within unchanged total budgets: 1,536 fit, 64 stop, 64 selection, and 256 internal requests/task. Report each task and the mixture. Source-trained BoolQ is not unseen-task transfer. T freezes its task action before its 512-request check. R fits on 512

request labels, freezes on 128 selection requests, and tests 512 internally, including every online cost and fallback. Its stronger fixed-task comparator is remeasured in the same window. These are separate scope-specific paths; their results cannot be pooled into a joint M/T/R deployment.

5 Results: Realizability, Identifiability, and Reusability

Bounded compensation realizes joint-positive removals across multiple backbones. The following comparisons then distinguish the selector's contribution and the conditions for reusing those gains.

5.1 Q1: Realizable gains under bounded compensation

Joint gains across fixed crossings. Independently fitted rank-96 bridges improve CE, accuracy, and time at 20 of 24 sites: five on Qwen3, seven on Qwen1.5, and eight on Qwen2.5 (Table 1). Each backbone has eight fixed-source spans (1–8) at equal per-site budgets. Each site's identical checkpoint is evaluated across two 512-request roles. Thus limited local compensation preserves or improves task capability while replacing original computation across generations and sizes.

Qwen3 spans 5/7/8 and Qwen1.5 span 7 save time but worsen CE and accuracy; span alone therefore cannot determine admission. Appendix D reports all 24 outcomes, rather than only their ranges.

Adaptation improves quality; replacement saves computation. The no-skip adapter improves quality more but adds latency; replacement improves both quality and time (Table 2). Thus quality recovery does not require removal, whereas the shorter path makes that recovery cheaper in this matched comparison. Replacement-minus-no-skip $\Delta U = +.476$ ($[-.337, 1.290]$) remains uncertain. The released MLP/MSE port saves time but fails selection quality and freezes F_0. Against this frozen fallback policy, the primary's paired ΔU is $+2.448$ ($[1.554, 3.343]$). Comparing the two raw replacements diagnoses capture: $\Delta U = +3.977$ ($[2.990, 4.964]$), CE difference $+.18116$, and accuracy difference $+10.938$ points, with no resolved speed advantage ($-.022$ ms, $[-.064, .020]$). This supports the primary recipe at this site; the differing ingredients do not identify a unique cause.

5.2 Q2: Identifying the executable opportunities

Automatic identification and its quality–time trade-off. Four policies nominate from the same nine source-crossfitted sites (Table 3). Executed-gain supervision freezes $13 \to 14$, teacher-quality supervision $10 \to 11$, and geometry $12 \to 15$; all pass the empirical gate. Executed-minus-teacher $\Delta U = +.489$ ($[-.457, 1.434]$); executed-minus-geometry $-.784$ ($[-1.663, .094]$). Native importance chooses the same action as executed-gain supervision. The predictor therefore identifies a useful task replacement, but does not establish superior selection under the fixed composite objective. Its CE advantage over geometry is $+.08300$ ($[.05197, .11403]$), offset by smaller time savings. This answers Q2 partially: automatic identification is feasible in this pool, whereas comparative utility advantage remains unresolved.

5.3 Q3: Reusing gains across backbones and decision horizons

The task-fixed workflow transfers across model families. The same bridge structure, training rule, and select–fit–measure–freeze–check procedure produce three-mean-positive actions on Qwen3, SmolLM2, and Pythia (Table 4). Together with Table 1's within-Qwen breadth, this is cross-backbone evidence beyond the fixed-site controls. Geometric nomination uses midpoint-minus-two offsets $(+1, +3, +5)$; 128 selection requests precede 512 internal requests. This development revision followed an earlier span-prior run that refused every Pythia candidate; that failure remains. Transferability here means the mechanism and training procedure adapt with model-specific weights, not direct checkpoint or learned-selector transfer.

Table 1: Separately post-trained rank-96 bridge checks versus paired F_0; ranges include all eight predeclared crossings per model, including negative quality cells. Time is saved milliseconds per pretokenized batch-1 no-cache model path; accuracy is percentage points. Each site has 1,024 paired requests across two historically opened roles.

Backbone	CE gain range	Accuracy gain range	Time saving range
Qwen3-0.6B	−.113–+.184	−7.62–+9.77	+.774–+8.341
Qwen1.5-1.8B	−.021–+.307	−1.95–+10.45	+1.232–+10.106
Qwen2.5-1.5B	+.123–+.298	+3.52–+10.25	+1.042–+8.968

Table 2: Fixed $13 \to 14$ Qwen3/HellaSwag controls, 512 paired internal requests per block. No-skip matches architecture, teacher, and fitting budget; MLP/MSE matches data and student updates, not parameters or teacher objectives. Each block re-times both methods; blocks are separate windows. Accuracy is points, net time milliseconds. The raw port failed selection quality and froze F_0.

Execution	CE gain	Acc. gain	Net time	U	Emp. gate
Bridge replacement	+.09796	+3.320	+.827	2.471	pass
Original layers + adapter	+.13202	+5.078	−.342	1.994	fail
Bridge (port-comparison window)	+.09796	+3.320	+.811	2.455	pass
Released MLP/MSE port, raw	−.08320	−7.617	+.833	−1.522	fail
Port's frozen F_0 policy	0	0	+.006	.006	fallback

The selected actions reduce paired F0 model-path times by 9.54%, 12.01%, and 3.51%, respectively (absolute baselines in Appendix D). Pythia's descriptive 95% intervals are CE $[+.05869, +.10139]$, accuracy $[+.151, +7.271]$ points, and time $[+.883, +.937]$ ms; Qwen3/SmolLM2 accuracy intervals include zero. Every +5 candidate is faster but fails internal quality: broader removal does not monotonically improve realizable benefit.

A shared checkpoint retains aggregate benefit, not uniform task benefit. Both M heads freeze the same $12 \to 15$ checkpoint. On the declared mixture, CE improves by +.04898, time saving is +2.921 ms, and accuracy change is zero ($U = 3.411$). These observed sample means show component-wise nonnegative aggregate improvement, although the original strict accuracy-improvement gate fails. Task cells reveal the reuse boundary: HellaSwag loses quality, while BoolQ gains it. A FIT/STOP diagnosis also shows incomplete teacher-to-bridge retention. On 64 HellaSwag STOP requests, the exact teacher gives CE +.05924/accuracy +6.250 points, versus student +.00406/0; BoolQ gives teacher +.31726/ + 15.625 versus student +.27322/ + 9.375. These diagnostic differences do not uniquely explain the task cancellation, but show why a teacher's positive response cannot substitute for measuring the fitted action.

Whole-model-held-out T prediction also failed: over 39 FIT sites/five backbones, both heads recover the admitted best on four backbones, but feature selection returns F_0 on Qwen3 and geometry fails internal quality ($\Delta U = -2.831$). Thus transferring the bridge-training procedure and transferring a learned selector are different empirical outcomes.

Request-conditioned reuse survives its online cost. After 512 FIT and 128 selection requests, the frozen R head improves all three internal means while paying for source observation, route selection, and one bridge/F_0 path (Table 3). Its three installed bridges are $10 \to 11$, $13 \to 14$, and $13 \to 16$, drawn from the fitted task library. Route counts are 146/125/176 with 65 fallbacks; fallback scores equal F0.

Against qualified fixed $13 \to 14$, paired $\Delta U = -.695$ ($[-1.593, .204]$); both policies pass. R saves slightly more time in this window but yields smaller quality gains. On 32 FIT requests, it matches measured-pool utility-best 12 times, with mean regret 2.544. The controller costs .561 ms; full R minus its identical forced action costs .583 ms ($[.315, .851]$), with equal scores. Conditional reuse is therefore feasible after recognition cost, but its incremental value over

Table 3: Qwen3 paths versus paired F_0, 512 internal requests/run. M uses one checkpoint on 50:50 HellaSwag/BoolQ; T/R use HellaSwag. Accuracy is points, time saved milliseconds. Empirical gates require three strictly positive means; M/T nomination is offline and R cost is included. Comparisons are paired within blocks only. The R block uses the qualified fixed-task control, not the earlier quality-negative control.

Scope / policy	Frozen action	CE	Acc.	Time	U	Emp. gate
M neutral / geometry	12→15	+.04898	0	+2.921	3.411	fail
HellaSwag cell	same weights	−.04207	−3.125	+2.917	1.871	fail
BoolQ cell	same weights	+.14002	+3.125	+2.926	4.951	pass
T executed-gain head	13→14	+.09796	+3.320	+.724	2.368	pass
T teacher-quality head	10→11	+.08395	+1.367	+.766	1.879	pass
T geometry head	12→15	+.01496	+.586	+2.885	3.152	pass
T native importance	13→14	+.09796	+3.320	+.724	2.368	pass
R joint controller	bank / F_0	+.04502	+2.148	+.810	1.690	pass
R-block qualified T	13→14	+.09796	+3.320	+.741	2.384	pass

Table 4: Geometric Top-3 on HellaSwag: selection-frozen actions evaluated on 512 internal requests against paired F_0. Accuracy is points; time is saved batch-1 no-cache model-path milliseconds, including the bridge. All three mean gates pass. U evaluates the frozen action, not internal-best selection or a cross-backbone ranking.

Backbone	Frozen action	CE gain	Accuracy gain	Time saving	U
Qwen3-0.6B	12→15	+.01496	+.586	+2.939	3.205
SmolLM2-1.7B	10→13	+.14818	+1.172	+4.463	6.179
Pythia-1.4B	10→11	+.08004	+3.711	+.910	2.452

task-fixed reuse is unestablished. The FIT oracle diagnoses choice error; it is not a deployable policy.

Appendix B retains the earlier non-primary request failure separately; it is not pooled with this bank.

6 Discussion and Limitations

Redundancy depends on what can be recovered cheaply. The results support operational redundancy under a bounded compensation procedure. An interval can have substantial hidden-state displacement yet admit a useful trained replacement. Conversely, a promising teacher does not ensure that a smaller student retains its benefit. Dynamic H addresses this distinction by learning the installed student's quality–time response and measuring the same action before reuse. Replacement modules and frozen-backbone post-training have precedents [2]; the contribution is their coupling to compensation-conditioned identification and scope-specific reuse, not the claim that either ingredient is new.

The matched no-skip control clarifies the empirical mechanism. Adaptation improves quality even when all original layers remain, so quality gains cannot be attributed to deleting intrinsically useless blocks. Replacement combines recovery with shorter execution. Its contrast with the released MLP/MSE port shows a better fitted response at one site, but differing architecture and supervision prevent attributing that difference to a single design choice. The evidence establishes useful compensated removal, rather than the necessity of every bridge ingredient.

Recoverability and identifiability are separate results. The bridges provide stronger evidence than the selectors. Automatic task nomination finds an action with joint gains and stronger CE improvement than geometry, but geometry has higher mean composite utility and native importance selects the same action. Task-effect features permit a broader decision interface; this expressive capacity is not proof of superior prediction. Similarly, positive

request execution demonstrates conditional reuse after online cost, but not an advantage over qualified task-fixed reuse. The identical-action control isolates a real recognition cost of .583 ms. Finer decisions are worthwhile only when their additional recoverable benefit compensates for that cost.

Procedure transfer and checkpoint reuse have different boundaries. Positive fixed-site Qwen outcomes and selection-frozen Qwen/SmolLM2/Pythia actions show that one bridge structure and training rule can adapt across backbones. This is substantive procedural transfer, not transfer of identical weights or a universal selector. Shared-checkpoint model reuse is a different test: mixture CE and runtime improve at unchanged accuracy, but opposing task effects fail the strict empirical gate. Aggregate benefit can therefore coexist with insufficient task-wide retention. The teacher–student diagnosis identifies a possible capture gap without proving it is the sole cause.

Limits of the evidence. Automatic nomination covers nine measured anchor sites; full-depth enumeration is not a validated global search. Historically inspected development roles and one seed limit independent generalization and training robustness. Positive sample means do not certify positive population means. BoolQ belongs to M's training distribution, and M/T/R are evaluated separately rather than as a joint controller. Segment controls are not official full-system reproductions [2, 5–7]. Timing concerns pretokenized BF16 batch-1 no-cache backbone paths, including bridges and online R decisions but excluding score observation symmetrically. Cached serving, energy consumption, and complete service latency are not measured. Within this setting, compensated removal is realizable; broader identification advantage and reliable cross-task retention remain open.

7 Conclusion

We study which Transformer computation can be removed under limited local compensation, and how its gains can be identified and reused. Dynamic H pairs fitted-action gain prediction with task-effective rank-96 replacement. Joint-positive crossings, matched adaptation controls, and cross-family procedure transfer show that bounded compensation can produce faster paths with better observed task quality. Automatic task and request policies provide usable instances; shared-checkpoint model reuse retains aggregate CE and time benefit but not strict three-component qualification. The central finding is that removable computation must be assessed through the replacement that realizes it. Bridge recoverability, selector quality, and reuse scope are distinct requirements: success in one does not establish the others.

Reproducibility Statement

Sections 2–4 define execution, features, compensation, and empirical admission. Appendix A.1 specifies the teacher, scoring, losses, and optimization; Appendix B gives uncertainty; Appendix C details controls and port adaptations; Appendix D reports all fixed crossings and retained absolute baselines. Per-site bridge records are under experiments/h_dynamic_v3_20260925/multispan_shared_v1; M/T/R code, checkpoints, identities, paired arrays, and recomputation scripts are under experiments/h_dynamic_v3_20260925/three_level_closure. Archive coverage differs: the two earlier Qwen local-reference series retain summary means rather than complete raw-array evidence. Upstream weights and datasets retain their original terms.

Ethics Statement

This work studies task-conditioned structural adaptation of public language-model checkpoints on public benchmarks. It uses no human-subject or private data and releases no new foundation model. The principal risk is silent quality degradation if a learned omission is reused outside its declared backbone, task, data distribution, or serving regime. Exact-F_0 refusal, model–task-bound manifests, and explicit reporting of negative quality or latency cells are intended to reduce this risk. Independent external testing is not established for this version, and none of these measures proves safety under arbitrary distribution shift.

AI Use Statement

Large-language-model tools, including OpenAI Codex, assisted code inspection, experiment orchestration, consistency checking, and manuscript editing. Numerical claims are checked against retained evaluator outputs and frozen artifacts rather than generated text. The authors remain responsible for the mechanism, experimental design, execution, citations, and final claims.

References

[1] Xin Men, Mingyu Xu, Qingyu Zhang, Qianhao Yuan, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. ShortGPT: Layers in Large Language Models are More Redundant Than You Expect. In Findings of the Association for Computational Linguistics: ACL 2025, pages 20192–20204, 2025. doi: 10.18653/v1/2025.findings-acl.1035. https://aclanthology.org/2025.findings-acl.1035/.

[2] Xiaodong Chen, Yuxuan Hu, Jing Zhang, Yanling Wang, Cuiping Li, and Hong Chen. Streamlining Redundant Layers to Compress Large Language Models. In International Conference on Learning Representations, 2025. https://openreview.net/forum?id=IC5RJvRoMp.

[3] Fei Wang, Li Shen, Liang Ding, Chao Xue, Ye Liu, and Changxing Ding. Layer as Puzzle Pieces: Compressing Large Language Models through Layer Concatenation. In Advances in Neural Information Processing Systems, volume 38, pages 6791–6824, 2025. doi: 10.52202/085713-0235. https://proceedings.neurips.cc/paper_files/paper/2025/hash/0a4096f09ab230082df548f901d4f79e-Abstract-Conference.html.

[4] Xuan Ding, Pengyu Tong, Ranjie Duan, Yunjian Zhang, Rui Sun, and Yao Zhu. Pruning as a Cooperative Game: Surrogate-Assisted Layer Contribution Estimation for Large Language Models. In International Conference on Learning Representations, 2026. https://openreview.net/forum?id=Azgxj7ZGCX.

[5] Paweł Batorski, Abtin Pourhadi, Akylgali Aitaza, Przemysław Spurek, and Paul Swoboda. MACRO: Markov Chain Routing of Transformer Layers. arXiv preprint arXiv:2608.05872, 2026. doi: 10.48550/arXiv.2608.05872. https://arxiv.org/abs/2608.05872.

[6] Akriti Jain, Saransh Sharma, Koyel Mukherjee, and Soumyabrata Pal. FiRST: Finetuning Router-Selective Transformers for Input-Adaptive Latency Reduction. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21957–21975, 2025. doi: 10.18653/v1/2025.findings-emnlp.1197. https://aclanthology.org/2025.findings-emnlp.1197/.

[7] Ahmed Heakl, Martin Gubri, Salman Khan, Sangdoo Yun, and Seong Joon Oh. Dr.LLM: Dynamic Layer Routing in LLMs. In International Conference on Learning Representations, 2026. https://openreview.net/forum?id=3ow7tq0O3l.

[8] Minkyu Kim, Vincent-Daniel Yun, Youngrae Kim, Suin Cho, Woosang Lim, and Sunwoo Lee. Rethinking Layer Redundancy: Calibration Matters More Than Search in LLM Depth Pruning. arXiv preprint arXiv:2604.24938, 2026. doi: 10.48550/arXiv.2604.24938. https://arxiv.org/abs/2604.24938.

A Implementation and Evaluation Details

The evaluated D-Module has independently post-trained rank-96 weights at each crossing. Each action records its teacher companion, student checkpoint, data roles, and timer. Shared-trunk or rank-8-adapter variants are different interventions and are not pooled with this bridge.

Table 5: Executable mapping of the Dynamic-H fields. These invariants keep recognition and removal aligned; they are not separate novelty or performance claims.

Field	Operation	Required distinction
G_{meta}	Legal endpoint enumeration linked to model, site, bridge and scope	incomplete coverage or invalid state/cache geometry
ρ	Model structure, offline task effect, legal request-prefix features	task-specific input in M or post-decision information in R
Θ	Gain prediction, Top 3 nomination and fixed composite objective	nomination regret or post-test ranking changes
$\mathcal{M}$	Site-matched teacher and independently fitted light bridge	teacher positive but executable bridge negative
κ	One prefix–bridge–suffix path or exact F_0	hidden branch work or timer mismatch
$\mathcal{C}$	Fit/stop/selection/internal/external roles and admission	leakage, reselection or negative mean admitted

Scope-specific training. T's bridge fits on 3,072 HellaSwag requests with 128 stop requests. M uses 1,536 fit and 64 stop requests per task, equally weighting HellaSwag and BoolQ. The teacher's four stages have 768/768/384/384 total steps; M balances tasks within those steps rather than doubling them. Each site-specific rank-96 student has 256 hidden-distillation and 512 task steps. Task-step checkpoints 0/128/256/384/512 compete only by stop-role CE. For each source s, teachers use destination pairs $(s+1, s+5)$ and $(s+1, s+3)$; the short student is fitted once with the first pair. Candidate students do not share weights. R fits a fixed ridge head on 512 request-level counterfactual labels from the exact task bridge bank, not on M weights or teacher gains.

Table 6: Identity positions for the new checks; intervals are half-open. HellaSwag positions index its frozen train manifest; BoolQ positions index eligible source rows after the fixed 192-token length check. Original dataset row IDs are retained in locks and raw arrays. Neither M source task is an unseen external task.

Protocol / task	Bridge fit / stop	Selection	Internal
M / HellaSwag	$[0, 1536)$ / $[3072, 3136)$	$[3200, 3264)$	$[3328, 3584)$
M / BoolQ	$[0, 1536)$ / $[1536, 1600)$	$[1600, 1664)$	$[1664, 1920)$
Automatic T / HellaSwag	$[0, 3072)$ / $[3072, 3200)$	$[3200, 3328)$	$[3328, 3840)$

Feature/label and prediction checks. M's unlabeled integer-token probe has seed 2026092604, 16 sequences of 64 tokens, and 378 endpoint feature rows; boundary/native F0 drift is zero. Its 18D input has 14 structural/probe fields plus four geometry fields, no task gradients. T's 16D head uses compute geometry, bridge-state summaries, task-effect summaries, and three interactions; its six-field geometry comparator includes teacher companion coordinates. FIT gain labels use 32 requests/site/task as appropriate. Sources 10/12/13 and spans 1/3/5 are retained. Each held-source head standardizes the other six sites, uses ridge $\alpha = 10$, and predicts that source's gains. All nine crossfitted predictions are ranked together; no favorable source is selected. Each policy measures at most three candidates and freezes its selection-role winner/F_0 before internal tests. Three-site held pools do not establish global Top-3 recovery.

For T, feature nominees are $10 \to 11$, $13 \to 14$, $12 \to 13$; geometry nominees are $12 \to 13$, $12 \to 15$, $13 \to 14$. M neutral and geometry nominate the same set, $12 \to 15$, $13 \to 16$, $12 \to 13$, in different orders. These lists are fixed before selection measurements, not reconstructed from internal-best actions.

Observation and uncertainty. Native F0 outputs match the split executor for four-choice HellaSwag and two-choice BoolQ. Automatic comparisons use same-device F0–action–F0

brackets, alternating action order, four warmups, BF16/eager, batch 1, and no cache. Timing ends at the final backbone hidden state; task-score observation follows on both sides. M averages within-task means with weights $w_\tau = .5$; the estimated variance of its mean is $\sum_\tau w_\tau^2 s_\tau^2/n_\tau$. R times its once-per-request features, head, dispatch, and chosen path, including fallback.

Paired arrays, row identities, checkpoints, complete hashes, and scripts are retained under experiments/h_dynamic_v3_20260925/three_level_closure. They support recomputation of means, intervals, frozen choices, and gates. Within-model and held-model selector tests have separate records. Within-run role separation does not make historically inspected data independent external evidence.

A.1 Task scoring and post-training objectives

For option c with continuation positions I_c, the original LM head and final normalization produce

$$v_c(x) = \frac{1}{|I_c|}\sum_{t\in I_c} \log p(x_{c,t+1} \mid x_{c,\le t}), \qquad L_\tau(v,y) = -v_y + \log\sum_c e^{v_c}, \qquad \hat{y} = \arg\max_c v_c. \tag{9}$$

HellaSwag uses four options; BoolQ uses the two supplied answer forms. Padding and prompt-only positions do not contribute to v_c. The final-score observer uses this same definition for F0, teachers, and students. Subtracting the mean option score leaves CE and the decision unchanged. One request evaluates all supplied options; R may use their prefix states, not the correct-option label.

Teacher architecture. Let r and d_{long} denote the short and long destinations, with $s < r < d_{\text{long}}$. The source-conditioned direct branch is

$$u_j = \mathcal{X}_j(A_j(T_j(A_s(h_s))),\, \mathcal{S}_s(h_s))\,, \qquad j \in \{r, d_{\text{long}}\}. \tag{10}$$

Each A is a residual LayerNorm–$H \to 96 \to H$ GELU adapter with learned scale initialized at .1; T_j is a residual $H \to 96 \to H$ GELU transition. $\mathcal{S}$ is causal residual self-attention and $\mathcal{X}$ residual cross-attention. Both use four heads with total attention width 256, rotary positions, a width-256 SiLU MLP, LayerNorm, padding masks, and zero-initialized residual output projections. The hierarchical short state is $\tilde{u}_r = \mathcal{S}_r(u_r)$; the long state is $\tilde{u}_{d_{\text{long}}} = \mathcal{X}_{\text{ext}}(u_{d_{\text{long}}}, \tilde{u}_r)$. The deployed student distils u_r for the short destination or $\tilde{u}_{d_{\text{long}}}$ for the long destination. No teacher executes during deployment.

Structural objectives. Let $R(u,v)$ average masked per-token squared error divided by $\max(H^{-1}\|v_t\|^2, 10^{-5})$. Let $D(u,v)$ be masked mean squared error after separately normalizing each state by $\sqrt{H^{-1}\|u_t\|^2 + 10^{-6}}$. For the unchanged destination block f_j, define

$$K_j(u) = R(u,h_j) + .5D(u,h_j) + R(f_j(u),h_{j+1}) + .5D(f_j(u),h_{j+1}), \tag{11}$$

$$S_j(u) = K_j(u) + .25R(u-h_s, h_j - h_s). \tag{12}$$

The base structural loss is $[S_r(u_r) + 1.5S_{d_{\text{long}}}(u_{d_{\text{long}}})]/2.5$. Its task stage adds the corresponding weighted task losses. Hierarchical reconstruction uses

$$S_{\text{hier}} = .5K_r(\tilde{u}_r) + K_{d_{\text{long}}}(\tilde{u}_{d_{\text{long}}}) + .25R(\tilde{u}_{d_{\text{long}}} - h_s, h_{d_{\text{long}}} - h_s). \tag{13}$$

Base stages train the direct branches; hierarchical stages freeze them and train only the midpoint context and long extension.

Teacher task objective and schedule. For candidate scores v and F0 scores v_0, with centered scores v°, the task loss is

$$P(v) = L_\tau(v,y) + .1\,\text{MSE}(v^\circ, v_0^\circ) + 2(.05)\,\text{softplus}\left(\frac{L_\tau(v,y) - L_\tau(v_0,y) + .01}{.05}\right). \tag{14}$$

The hierarchical task stage minimizes $.5S_{\text{hier}} + P(v_{\text{long}})$. Both task stages add .01 times the mean parameterwise squared distance from their structural-stage initialization. AdamW uses weight decay .001, gradient clipping at norm 1, and stage learning rates $(2\cdot 10^{-4}, 5\cdot$

$10^{-5}, 2\cdot10^{-4}, 5\cdot10^{-5}$) for $(768, 768, 384, 384)$ updates. Every 192 updates, the lowest stop-role stage objective is retained for the next stage. Task balancing in M preserves these total budgets.

Student objective. Write $u = B_{s,d}(h_s)$ and let $\tilde{h}_d, v_{\text{teacher}}$ be the frozen teacher's destination and final scores. The student hidden error is

$$E_h(u, \tilde{h}_d) = \frac{\sum_t m_t \, \|u_t - \tilde{h}_{d,t}\|^2/H \,/\left(\|\tilde{h}_{d,t}\|^2/H + 10^{-5}\right)}{\sum_t m_t}. \tag{15}$$

It differs slightly from the teacher's clamped normalization. Hidden distillation minimizes E_h for 256 updates at learning rate $3\cdot10^{-4}$. The next 512 updates minimize $L_\tau(v, y) + .2\,\text{MSE}(v^\circ, v^\circ_{\text{teacher}}) + .05E_h$ at $5\cdot10^{-5}$. Both stages use AdamW with weight decay .001 and norm-1 clipping. Stop CE selects among task updates $0, 128, 256, 384, 512$; ties retain the earlier checkpoint. The frozen suffix transmits gradients but receives no parameter updates.

A.2 Local sensitivity and nomination limits

Equation (7) concerns the differentiable, real-valued counterpart of the fixed-mask evaluator. Taylor's integral remainder gives $r_a = \int_0^1 (1-t)\delta_a^\top \nabla^2 \ell_x(h_d + t\delta_a)\delta_a \, dt$. If the Hessian operator norm on this segment is at most β_x, then $|r_a| \le \beta_x \|\delta_a\|^2/2$. Thus a correction aligned against the task-loss gradient can lower loss, while a small norm alone supplies no sign guarantee. The expansion explains the training target; β_x is not estimated, and the bound is not used to certify measured gains. Accuracy is discontinuous at option-score ties and must be observed separately.

Prediction accuracy also limits nomination. In a finite pool of at least four actions, suppose each predicted gain component has uniform error at most ϵ_j. Then scaled utility error is at most $\epsilon_U = \sum_j \epsilon_j/q_j$, with $q = (.1, .05, 1)$ in CE/fraction/ms units. A true utility-best action is predicted admissible if each of its gains exceeds ϵ_j. If it also exceeds the fourth-best true utility by more than $2\epsilon_U$, it must enter predicted Top 3: no more than two rivals can outrank it. This conditional argument does not guarantee global recovery; neither uniform error bounds nor full-depth gain labels are established by the nine-site experiment. More request labels at a site do not increase the number of independent endpoint examples.

B Uncertainty and Matched Comparison Details

Intervals below are descriptive paired 95% estimates ($\bar{g} \pm 1.96\,\text{SE}$), not multiplicity-adjusted or multi-seed guarantees. A positive mean passes even when its interval includes zero; an exact zero mean does not pass. Admission and uncertainty are reported separately.

Table 7: Internal component uncertainty for frozen Qwen3 policies. T/R rows retain their earlier windows; Table 3 uses the four-policy T and qualified-control R windows. M uses stratified task-mixture variance; T/R use paired request variance. Accuracy intervals are points; time intervals milliseconds. Scope rows are not paired head-to-head comparisons.

Policy	CE 95% interval	Accuracy 95% interval	Time 95% interval	U 95% interval
M neutral/geometry	$[.00082, .09713]$	$[-3.463, 3.463]$	$[2.906, 2.937]$	$[2.409, 4.413]$
T task-feature	$[.06763, .12829]$	$[-.324, 6.964]$	$[.755, .820]$	$[1.540, 3.323]$
R joint controller	$[.01018, .07985]$	$[-1.075, 5.372]$	$[.544, .774]$	$[.687, 2.391]$

Matched utility comparisons. M selectors share the same checkpoint, so paired $\Delta U = 0$ with interval $[0, 0]$. Earlier within-model T feature minus geometry gives $-.763945$ ($[-1.646288, .118398]$); its original timing window remains separate. R minus fixed $13 \to 16$ gives -1.270169 ($[-2.146274, -.394064]$); negative quality excludes that control despite higher raw U. The subsequent same-window qualified $13 \to 14$ comparison gives fixed T CE $+.097960$, accuracy $+3.320313$ points, savings $+.740767$ ms, $U = 2.384426$; unchanged R gives $+.045016/+2.148438/+.809896$, $U = 1.689745$. Both pass; R-minus-T $\Delta U = -.694681$

([−1.593265, .203903]) does not establish routing advantage. No controller refit or cross-window pooling occurs. These are segment comparisons, not official full-system reproductions.

Stronger transfer check retained. Whole-model-held-out T prediction uses 39 FIT sites across five backbones, excluding all Qwen3 sites for that target. Feature nominees are $13 \to 18$, $12 \to 17$, $10 \to 15$; geometry nominees are $13 \to 18$, $12 \to 17$, $13 \to 16$. Feature selection returns F_0; geometry freezes $13 \to 16$. Its 512 internal means are CE $-.001346$, accuracy $-.585938$ points, and time $+2.961706$ ms; both fail admission. Paired $\Delta U = -2.831059$ ($[-3.845248, -1.816870]$). This failure remains separate from within-model adaptation, whose head legitimately uses other-source Qwen3 FIT labels.

Matched no-skip protocol and uncertainty. The frozen replacement is $13 \to 14$, with checkpoint hash beginning 2d62d6ddae7a; its exact teacher uses 13/14/18. For source h_s and original destination $h_d = F_{s:d}(h_s)$, the no-skip control supplies $h_d + (B_{\text{control}}(h_s) - h_s)$ to the same suffix. Its rank-96 source residual has the same parameter count, teacher access, seed 1, 3,072 FIT and 128 stop requests, optimizers, losses, and 256 hidden plus 512 task steps. All original blocks execute. Zero up-projection gives exactly F0; native and zero-adapter final-logit smoke drift are zero. Stop CE chooses step 512. The control is separately trained, not the primary bridge added to original layers; fitting cost and initialization match, but the fitted weights differ.

The fixed site receives 128 selection-role observations and then 512 internal observations, with no new site or policy selection. Timer brackets remeasure both policies on GPU RTX 5060 Ti in the same BF16/eager batch-1 no-cache window, excluding the score observer on both sides. Replacement descriptive 95% intervals are CE $[.067625, .128294]$, accuracy $[-.324, 6.964]$ points, time $[.747991, .905898]$ ms, and U $[1.581255, 3.359953]$. No-skip intervals are $[.093693, .170343]$, $[1.552, 8.604]$ points, $[-.428377, -.255012]$ ms, and $[1.077927, 2.910297]$. Replacement-minus-control differences are CE $-.034058$ ($[-.061524, -.006593]$), accuracy -1.758 points ($[-5.028, 1.513]$), time $+1.168639$ ms ($[1.057153, 1.280124]$), and $\Delta U = +.476492$ ($[-.337430, 1.290415]$). These paired differences do not establish composite superiority.

The matched_bridge_control_v1 bundle retains both role arrays and the fitted control. It is a fixed-site bridge comparison, not an official replacement-method reproduction.

C Matched Recognition, Capture, and Released-Port Checks

The four checks retain frozen bridge identities and strict mean gates. They use seed 1, RTX 5060 Ti devices, and historically opened development roles; intervals are descriptive paired estimates. No new sites, outcome-dependent thresholds, or independent external identities are introduced.

Executed-gain supervision versus controls. Each of the nine T checkpoints is observed on the same first 32 FIT requests, with its exact hierarchical teacher companion retained. The source-held-out ridge head uses $\alpha = 10$, unchanged legal task/endpoint features and utility scales $(.1, .05, 1)$ for CE, accuracy fraction, and milliseconds. Teacher supervision replaces only CE/accuracy labels with exact-teacher outcomes; cost remains the deployed student's measured cost. No teacher runtime is mislabeled as light-bridge cost. Native importance ranks ascending normalized original-interval hidden displacement. All four policies have at most three candidates, the same selection/internal requests, and one paired timing window.

Executed-gain nominees are $10 \to 11$, $13 \to 14$, $12 \to 13$; teacher-quality nominees $10 \to 15$, $10 \to 13$, $10 \to 11$; geometry $12 \to 13$, $12 \to 15$, $13 \to 14$; native importance $12 \to 13$, $13 \to 14$, $10 \to 11$. Selection freezes $13 \to 14$, $10 \to 11$, $12 \to 15$, and $13 \to 14$, respectively. Executed-minus-teacher $\Delta U = .488927$ ($[-.456616, 1.434469]$), executed-minus-geometry $-.784269$ ($[-1.662689, .094151]$), and executed-minus-native 0. Executed-minus-geometry CE .083003 ($[.051973, .114034]$) does not establish composite superiority. Native importance is an ablation, not a full official pruning algorithm.

Frozen R recognition and cost. On its first 32 FIT requests, the unchanged head's mean chosen utility is 6.898969, label-informed measured-pool best 9.442580, and regret 2.543611. The chosen route is utility-best on 12/32; three fallbacks have another positive-utility action, and two choices have negative measured utility. These are in-sample, per-request diagnostics, not external prediction accuracy or an installable oracle.

Cached-prefix feature extraction costs .390930 ms, features plus joint head .486668 ms, and controller including dispatch .560880 ms. Full R versus forcing its identical action adds .583083 ms ([.315342, .850824]); score drift is zero. Prefix/bridge/suffix stay timed on both sides while the experimental observer is excluded. This distinguishes online cost from measurement-only work; offline M/T discovery is not charged per request.

M teacher-to-student capture. The single source-mixture-trained $12 \to 15$ student is compared with its exact 12/13/15 teacher on 32 FIT and 64 STOP requests per task. On HellaSwag STOP, student-minus-teacher CE is $-.055178$ ($[-.110669, .000313]$), accuracy -6.250 points. On BoolQ STOP, CE is $-.044039$ ($[-.140871, .052792]$), accuracy -6.250 points. Teacher/student means are reported in Section 5. The teacher is diagnostic, not a cost-free deployable replacement. These cells expose a retention gap but do not establish a unique cause or repair the failed internal M gate.

Released Streamline MLP/MSE port and budget. We use the released alternative under examples/mlp_mseloss_train_without_memory_issue, repository https://github.com/RUCKBReasoning/LLM-Streamline, commit adf50baa83a1fabb644a2bd165fb9636838ff00b. The biased $h \to 4h \to h$ ReLU MLP has no residual or normalization and predicts frozen original h_{14} from original h_{13} by ordinary mean MSE, including collated padding. AdamW has learning rate $2 \cdot 10^{-4}$, weight decay .001, and betas (.9, .95); the released cosine scheduler uses 3% warmup and minimum rate $5 \cdot 10^{-6}$. Qwen3 initialization uses normal standard deviation .02 and zero biases. Float32 parameters use BF16 MLP computation. The original primary $13 \to 14$ checkpoint and teacher are not refitted.

The port adapts architecture binding to Qwen3, data to HellaSwag with maximum 192 tokens, and the budget to 768 optimizer updates with one-request microbatches and accumulation two. This matches primary student updates $256 + 512$, not total teacher-development cost, parameters, tokens, or GPU-hours. Its 8,393,728 trainable parameters exceed the primary's 296,960. Seven stop-MSE checks at updates $0, 128, \ldots, 768$ replace the released large-batch one-epoch/500-microstep validation schedule; lowest stop MSE selects update 768. Formal port fitting takes 61.996 seconds (.017221 GPU-hours), excluding acquisition, validation runs, and prior bridge/teacher fitting. These adaptations prevent calling this the flagship inherited-Transformer or a full official reproduction.

Native/split logits and scores, checkpoint roundtrip, and frozen-backbone gradients pass smoke. Execution traces show all 28 original blocks in F0 and all except block index 13 in both replacements, with exactly one respective replacement call. Data roles match Table 6's T row. On 128 selection requests, raw port CE $-.060438$, accuracy -8.593750 points, and saving $+.762834$ ms fail quality, freezing F0 before 512 internal requests. Table 2 reports both raw internal action and primary. F0 policy CE/accuracy are exactly zero; its .006190 ms time difference ($[-.013590, .025970]$) is measurement noise, not admission or acceleration.

Primary-minus-raw-port differences are CE .181163 ($[.147981, .214345]$), accuracy 10.937500 points ($[6.994676, 14.880324]$), time $-.022118$ ms ($[-.063999, .019762]$), and $\Delta U = 3.977010$ ($[2.989536, 4.964483]$). Primary-minus-frozen-F0-policy $\Delta U = 2.448409$ ($[1.553863, 3.342955]$). Absolute F0 CE/accuracy are 1.14928/53.125%. Its paired mean model-path times are 31.57691 ms for the primary and 31.57129 ms for the raw port; their executed times are 30.76597 and 30.73824 ms. Separate F0 brackets account for the small difference in baseline timings.

Reproducibility. The four comparison bundles, scores where archived, timing arrays, source identities, and audit_four_task_evidence.py are retained with the M/T/R implementation. Full hashes and procedural logs accompany the code rather than the scientific exposition. The fixed-site port comparison and selector/M capture diagnoses answer different questions.

Non-primary request failure retained. The earlier non-primary request route gives CE +.02989, saving +.568 ms, and accuracy −.39 points over 4,096 replayed requests. It fails the strict joint gate and is not pooled with the independently post-trained rank-96 bank.

D Complete Fixed-Site Outcomes and Absolute Baselines

Table 8 expands the ranges in Table 1. Each row concerns its own independently post-trained rank-96 checkpoint and pools two equal-size, 512-request development roles. It does not compare selectors or aggregate different checkpoints. Qwen3 values are drawn from the per-site local-reference fields of the archived summary; Qwen1.5/2.5 values are archived local-reference summaries transcribed to eight decimals. The latter retained records do not permit a complete new raw-array uncertainty audit, so we report their means without reconstructing unavailable intervals.

Table 8: All 24 fixed-source outcomes. CE is F0 minus bridge task CE; accuracy is bridge minus F0 in points; time is paired model-path saving in milliseconds. Composite utility uses Equation (8). A positive utility does not override a negative component.

Backbone	Crossing	CE gain	Acc. gain	Time saving	U	Emp. gate
Qwen3-0.6B	15→16	+0.18411	+9.766	+0.774	4.568	pass
Qwen3-0.6B	15→17	+0.17528	+9.375	+1.885	5.513	pass
Qwen3-0.6B	15→18	+0.08882	+6.348	+2.918	5.076	pass
Qwen3-0.6B	15→19	+0.04465	+2.051	+3.960	4.817	pass
Qwen3-0.6B	15→20	−0.05002	−3.516	+5.122	3.919	fail
Qwen3-0.6B	15→21	+0.00672	+2.051	+6.106	6.583	pass
Qwen3-0.6B	15→22	−0.06841	−4.102	+7.270	5.765	fail
Qwen3-0.6B	15→23	−0.11251	−7.617	+8.341	5.692	fail
Qwen1.5-1.8B	13→14	+0.30683	+10.449	+1.232	6.390	pass
Qwen1.5-1.8B	13→15	+0.26780	+8.887	+2.506	6.961	pass
Qwen1.5-1.8B	13→16	+0.16277	+5.957	+3.815	6.634	pass
Qwen1.5-1.8B	13→17	+0.18336	+3.613	+5.026	7.582	pass
Qwen1.5-1.8B	13→18	+0.13088	+3.125	+6.310	8.244	pass
Qwen1.5-1.8B	13→19	+0.07871	+0.977	+7.575	8.558	pass
Qwen1.5-1.8B	13→20	−0.02063	−1.953	+8.836	8.240	fail
Qwen1.5-1.8B	13→21	+0.08372	+0.879	+10.106	11.119	pass
Qwen2.5-1.5B	15→16	+0.29783	+10.254	+1.042	6.071	pass
Qwen2.5-1.5B	15→17	+0.26111	+7.227	+2.179	6.236	pass
Qwen2.5-1.5B	15→18	+0.28044	+8.008	+3.295	7.701	pass
Qwen2.5-1.5B	15→19	+0.26202	+6.641	+4.445	8.393	pass
Qwen2.5-1.5B	15→20	+0.29321	+7.813	+5.576	10.070	pass
Qwen2.5-1.5B	15→21	+0.19678	+4.883	+6.706	9.650	pass
Qwen2.5-1.5B	15→22	+0.19185	+4.102	+7.825	10.563	pass
Qwen2.5-1.5B	15→23	+0.12257	+3.516	+8.968	10.897	pass

Table 9: Absolute paired F0 and selected-action baselines for Table 4, from the same 512-request arrays. CE is option-level CE, accuracy is percent, and time is model-path milliseconds. The percentage is $(t_0 - t_a)/t_0$, not a reciprocal speedup ratio. Each timing belongs to its own model and measurement window.

Backbone	F0 CE	Action CE	F0 acc.	Action acc.	F0 time	Action time	Saved
Qwen3-0.6B	1.14928	1.13433	53.125	53.711	30.805	27.867	9.54%
SmolLM2-1.7B	.94829	.80011	69.141	70.313	37.175	32.712	12.01%
Pythia-1.4B	1.15689	1.07685	51.563	55.273	25.909	24.999	3.51%