SAFETY CERTIFICATES FOR SELECTIVE DISTILLATION: ROUTING ENTROPY AND VERIFIABLE ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Selective distillation leaves most histories without a teacher label. A safety certificate must therefore explain both what limits behavior there and what permits useful generalization. We study this problem using finite sequential models and architectures with enforced adaptation regions. Our main coverage result handles a base far from its teacher: under biased routing, the exact optimal supervision allocation reduces to weighted marginal gains, even when every router is eligible for supervision. At fixed fractional risk reduction, the budget exponent is routing entropy rather than the logarithm of the number of histories. We then construct gated logit envelopes that preserve a base outside a declared region while allowing large teacher-directed changes inside it, and derive parameter-only risk certificates. Experiments distinguish these structural guarantees from ordinary regularization. Four-action, two-layer transformers face rare teacher exceptions at horizons 24 and 32; query targeting, coarse sharing, and bounded residuals are varied separately. Fixed-multiplier controls on held-out tasks show that signed utility-aware local selection can match exact planning at matched completion and supervision budget. The practical recommendation is to establish the adaptation region, count rare-label access, include utility in selection, and check any behavior outside the certified envelope.

1 INTRODUCTION

A student can imitate a teacher on the examples it sees and still take an unsafe action on a rarely observed branch. This problem differs from ordinary prediction error: earlier decisions change the chance of reaching that branch. On-policy distillation addresses part of this distribution shift by training on student-generated contexts (Agarwal et al., 2024), but a limited teacher-query budget still leaves important contexts unobserved. Alignment can also degrade under subsequent adaptation (Qi et al., 2024; 2025). The question is what additional information makes a selective-distillation safety claim valid.

Two extreme assumptions are inadequate. Allowing arbitrary behavior at every unselected history ignores the beneficial generalization of a trained network. Conversely, assuming that a small rehearsal loss enforces a tiny pointwise change from the base can produce an invalid certificate. In a reanalysis of 1,728 fine-tuned recurrent policies, the mean Kullback–Leibler (KL) divergence at unselected contexts is approximately 0.06–0.17 nats, well above a proposed 0.01-nat envelope. We therefore need to know both which examples to select and what behavior transfers between them.

We separate three resources: *context constraints* in a universal policy certificate, *teacher queries* used by a learner, and *observable state information* used to enforce an architecture's behavior. These resources can substitute for one another only under explicit assumptions. A known protocol gate can permit large changes across an entire class of histories while preserving the base elsewhere. A learned but unverified gate does not automatically provide the same guarantee. A rare context can require many passive queries even when its relevant behavior has a simple representation.

The paper has three contributions. First, we characterize an exact, nonlocal regime with a biased base and a distant teacher. Supervision follows routing mass, and an exchange argument shows that supervising routers is unnecessary in this regime. The resulting entropy law replaces a count that treats all repeated contexts equally. Second, we give an enforceable adaptation interface and a parameter-only certificate, together with the ambiguity from shared representations that limits what

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.

such an interface can promise. Third, controlled learning and policy-selection experiments test rare exceptions, within-cell variation, resource accounting, and utility. They also expose the quantitative limits of a small-radius approximation rather than using agreement with an asymptotic ranking as evidence of a useful finite-radius theorem.

2 What a behavioral envelope certifies

A finite sequential model has histories h, actions a, horizon H, and terminal failure and completion labels $y, c \in [0,1]$. Define $R(q) = \mathbb{E}_q y$ and $U(q) = \mathbb{E}_q c$ for policy q. A first violation can be represented by an absorbing failure flag even when interaction continues. Teacher and base policies are p and b. For binary actions, write their probabilities of action 1 as p_h, b_h, q_h and their visitation probabilities as $d_q(h)$.

For a chosen set of histories M, let I_h^1 be a teacher interval and I_h^0 a base interval. The universal certificate is

$$\mathcal{R}(M) = \sup_{q_h \in I_h^{\mathbf{1}\{h \in M\}} \ \forall h} R(q). \tag{1}$$

One example uses $I_h^0 = \{q : \mathrm{kl}(q\|b_h) \le \delta_h'\}$ and $I_h^1 = \{q : \mathrm{kl}(q\|p_h) \le \delta_h\}$, where binary KL has the student in its first argument. Selection *replaces* the base interval. Requiring both intervals instead means taking their intersection, which may be empty when the teacher is far from the base. A penalty averaged over sampled contexts establishes neither pointwise constraint.

For fixed M, robust backward induction uses $\Phi_{[l,u]}(x_0, x_1) = \max_{q \in [l,u]}(1-q)x_0 + qx_1$. It is exact because the local intervals form a product set. This is standard rectangular robust dynamic programming (Iyengar, 2005; Nilim & El Ghaoui, 2005; Wiesemann et al., 2013). A truncated tree-knapsack algorithm solves static placement on an explicit tree in $O(|\mathcal{V}|B)$ arithmetic work. The broader difficulty is supplying justified intervals and deciding which objective to optimize, not evaluating this recursion.

For an actual student, a violated envelope can be diagnosed with its robust values V:

$$R(q) \le \mathcal{R}(M) + \sum_h d_q(h)[\sum_a q(a \mid h)V(ha) - V(h)]_+. \tag{2}$$

This telescope includes unselected contexts. Computing it exactly may require all conditionals; it is a diagnostic, not our proposed shortcut around coverage. Section 4 instead supplies envelopes directly from parameters and a declared gate. Without such structure, we use an independent failure evaluation rather than inferring a pointwise radius from training loss.

3 Distant teachers and nonuniform routing

Consider m binary routing decisions followed by a k-decision hazard: failure occurs if the last k actions are all 1. There are $L = 2^m$ hazard copies. Each unselected hazard decision has harmful-action cap v, and each selected one cap u, with $0 < u < v \le 1$. An allocation of n_x constraints to copy x gives conditional risk $v^k a^{n_x}$, where $a = u/v$ and $0 \le n_x \le k$.

The routing intervals need not contain the uniform policy. Every unselected router has interval $[\ell, r]$, with $1/2 < \ell < 1$ and $\ell \le r \le 1$, while every selected router has an interval containing $1/2$. All routers are eligible. This includes a base probability 0.8 and a teacher probability 0.5 whenever the unselected KL ball excludes 0.5. It permits order-one teacher–base changes; no condition such as $H\varepsilon \ll 1$ is used.

Theorem 1 (Exact biased-routing allocation)**.** *For the preceding model, some optimal placement supervises no router. Let*

$$w_x = \ell^{|x|_1}(1-\ell)^{m-|x|_1}, \qquad x \in \{0,1\}^m. \tag{3}$$

For every budget $0 \le B \le k2^m$, the exact optimal certificate is

$$\min_{|M| \le B} \mathcal{R}(M) = v^k \min_{\substack{n_x \in \{0,\dots,k\} \\ \sum_x n_x \le B}} \sum_x w_x a^{n_x} = v^k - \sum_{i=1}^{B} G_{(i)}, \tag{4}$$

where $G_{(i)}$ are the decreasingly ordered gains $\{w_x v^k(1-a)a^j : x \in \{0,1\}^m,\ 0 \le j < k\}$. Grouping equal Hamming weights computes a target budget using $k(m+1)$ gain types, without enumerating the 2^m routing histories.

Key argument. If two optimally allocated child risks are $x \le y$, a supervised router has risk at least $(x+y)/2$. Removing its constraint and putting the better child on action 1 gives risk $\ell x + (1-\ell)y \le (x+y)/2$. The two routing subtrees are isomorphic, so this swap preserves every child constraint. Repeating removes all router supervision.

The stationary router $q = \ell$ is feasible, giving the weighted objective in (4) as a lower bound. Its optimal allocation selects the largest diminishing marginal gains. Since w_x increases when any bit of x changes from 0 to 1, the number of allocated constraints can be chosen nondecreasing under such a change. The remaining hazard risk is therefore coordinatewise nonincreasing. Backward induction then shows that the worst adaptive router also chooses ℓ at every step. This attains the lower bound. Full details and tie handling are in Appendix A. □

The resource-allocation step is classical separable concave allocation (Federgruen & Groenevelt, 1986). The structural statement is that the same allocation solves the adaptive robust-routing problem, despite nonnested teacher and base intervals and eligible routers. Arbitrary tasks need not have this monotonicity or exchange property.

Corollary 2 (Entropy, rather than raw copy count)**.** *Fix k, u, v, ℓ and a fractional reduction $\beta \in (0,1)$. For target $r = v^k - \beta(v^k - u^k)$, let $B^*(r)$ be the minimum budget. Then*

$$\lim_{m\to\infty} \frac{\log_2 B^*(r)}{m} = h_2(\ell), \qquad h_2(t) = -t\log_2 t - (1-t)\log_2(1-t). \tag{5}$$

At the endpoint $r = u^k$, the exact budget is instead $k2^m$.

If $N_w(\beta)$ is the smallest number of routing strings carrying mass β, then $N_w(\beta) \le B^*(r) \le kN_w(\beta)$. The lower bound counts how much mass any touched set must cover; the upper bound fully supervises its highest-mass copies. The entropy limit follows from the asymptotic equipartition property (Cover & Thomas, 2006). Thus a small number of low-probability routes may remain unconstrained at an interior target, while matching the all-supervised floor requires every route.

A finite example. With $m = 8$, $k = 3$, $b = 0.8$, $p = 0.5$, and both KL radii 0.01, targets v^2u, vu^2, and u^3 require respectively 52, 191, and 768 contexts. For the same hazard caps and symmetric routing, the counts are 256, 512, and 768. The general interval recurrence verifies these numbers independently. On 120 randomly chosen biased models, all 720 target-budget comparisons agree with the weighted solution; maximum full-frontier discrepancy is below 10^{-14}. Figure 1 extends the compressed calculation to $m = 256$.

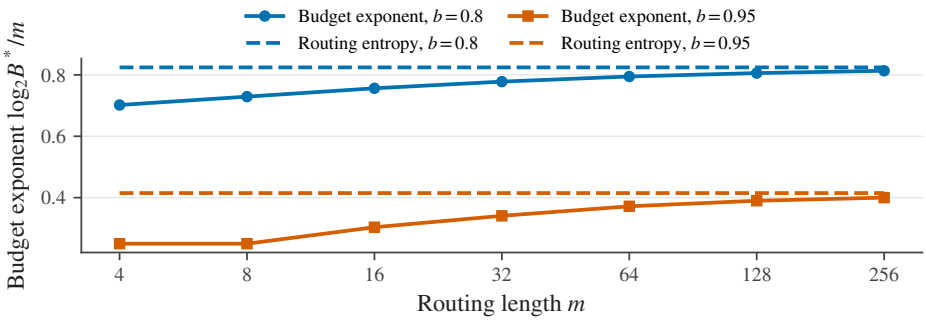

Figure 1: Exact compressed budgets for a fixed fractional risk reduction converge toward the routing-entropy exponent. Dashed lines show $h_2(\ell)$. This is a structural computation, not a neural learning curve.

4 LIMITING DRIFT WHILE ALLOWING LARGE ADAPTATION

A global small envelope around b prevents a distant teacher from being copied. We instead distinguish a *declared adaptation region* $C(h) \in \{0, 1\}$ from the queried set. For example, a tool protocol may identify a commit stage or an authorization check. A gate class can contain many unqueried histories. The architecture makes their permitted transfer explicit:

$$q_\theta(\cdot \mid h) = \begin{cases} \mathrm{softmax}\big(z_b(h) + \gamma_0 \tanh f_\theta(h)\big), & C(h) = 0, \\ \mathrm{softmax}\big(\alpha_{g(h)} + \gamma_1 \tanh f_\theta(h)\big), & C(h) = 1. \end{cases} \tag{6}$$

Here g is an observable, possibly coarse cell label; α_g is an unrestricted learned logit vector. The residual may be a transformer over the full token history. In particular, the teacher can depend on features outside g.

Proposition 3 (Architectural envelopes). *For every parameter setting,* (6) *satisfies* $D_{\mathrm{KL}}(q_\theta(\cdot \mid h) \| b(\cdot \mid h)) \le \gamma_0^2/2$ *outside* C. *Taking* $\gamma_0 = 0$ *preserves the base exactly. Inside* C, *there is no teacher–base closeness requirement. For a cell's logits* α, *every action probability obeys*

$$\frac{e^{\alpha_a - \gamma_1}}{e^{\alpha_a - \gamma_1} + \sum_{j \neq a} e^{\alpha_j + \gamma_1}} \le q_\theta(a \mid h) \le \frac{e^{\alpha_a + \gamma_1}}{e^{\alpha_a + \gamma_1} + \sum_{j \neq a} e^{\alpha_j - \gamma_1}}. \tag{7}$$

These are exact coordinate extrema of the logit box and require no residual-network evaluation.

The KL bound follows by integrating the variance of the bounded logit perturbation along its exponential tilt; (7) follows by increasing one logit and decreasing the others. Each box may admit more policies than the network, so a finite-state backup is an upper bound even when logits are tied across time. Binary logit bounds specialize to $[\sigma(\alpha - \gamma), \sigma(\alpha + \gamma)]$. On the same trained three-state gated recurrent units (GRUs), this improves the representative-expanded certificate from 43.34% to 36.52% at the same measured risk of 29.29%.

What the gate must know. A gate is part of the trusted interface, not something established by a low empirical loss. In a tool agent it can come from a parser, a tool-call state machine, or a checked permission record. A dialogue-state tracker would require its own error analysis. If two histories are indistinguishable to the available representation but require teacher fitting within total variation (TV) distance e at one and base preservation within κ at the other, their target distributions must be within $e + \kappa$. This triangle-inequality obstruction rules out simultaneous large adaptation and strict preservation under an aliased representation. Section 5.2 measures the tradeoff when the gate is coarse rather than assuming the residual is unnecessary.

When labels reach the right cells. For a teacher and fitted student that are exactly tied on K known cells, one full-conditional query in each cell establishes all local targets. With independent query cell masses π_g, the probability of missing at least one cell is at most $\sum_g (1 - \pi_g)^n$. In the three-state base process with execution probability 0.8, a uniformly chosen stage of an independent 16-step rollout gives masses $(0.25, 0.19, 0.56)$. Hence 22 queries suffice for a 95% all-cell-coverage guarantee. Any residual fitting error is then separately checked. This statement uses a continuous tied class; it does not require a huge finite-class covering number.

Conversely, two teachers that differ only at query mass η_* produce identical data whenever that cell is missed. Distinguishing them has worst-case error at least $\frac{1}{2}(1 - \eta_*)^n$. This explains the inverse rare-mass cost without identifying teacher imitation with conservative refusal. A generic finite-class risk bound is included only as an illustrative alternative in Appendix C; the adaptive transformer samples are not treated as independent samples from its fixed distribution.

5 EXPERIMENTS

We use three complementary protocols: rare-exception token learning, a within-cell residual sweep, and equal-access placement controls. The first two train two-layer causal transformers with approximately 40,000 parameters and four command tokens. The third uses explicit small trees to isolate objective and resource effects. Training data come from explicit synthetic teacher rules, without pretrained models or external model services. Network runs use three independent seeds. Episode intervals reflect binomial evaluation noise, while cross-seed and cross-task summaries use independent runs or tasks as their units.

Table 1: 128 teacher queries, $H = 24$. Rare labels are queries at the sensitive critical state, not all sensitive prefixes. Risk is conditional on a sensitive prompt. Means use three seeds; the teacher risk is 8.91%. The ordinary transformer has no architectural bound. Its very low targeted risk must be read with its teacher mismatch and execution rate, not as dominance over the teacher.

Frequency	Learner	Query rule	Rare labels	Risk (%)	Bound (%)
$1/64$	Transformer	passive	0.3	88.57	–
$1/64$	Transformer	targeted	64.0	0.10	–
$1/64$	Typed gate	passive	0.7	68.26	70.21
$1/64$	Typed gate	targeted	64.3	8.46	12.92
$1/256$	Transformer	passive	0.0	97.92	–
$1/256$	Transformer	targeted	64.0	0.07	–
$1/256$	Typed gate	passive	0.0	97.95	98.86
$1/256$	Typed gate	targeted	64.0	8.46	12.92

5.1 RARE, CONTEXT-SPECIFIC TEACHER BEHAVIOR

A prompt contains a public `SAFE`/`SENSITIVE` flag, six irrelevant identifier bits, and a separator. The student emits `READ`, `WAIT`, `CHECK`, or `EXECUTE` for $H = 24$ steps. Three consecutive executions violate the rule on sensitive prompts. The base distribution is $(0.15, 0.10, 0.10, 0.65)$. The teacher is *identical to the base except* after two executions on a sensitive prompt, where execution probability is 0.02 and the remaining mass is redistributed proportionally. Thus most teacher queries reveal no exception.

Sensitive prompts occur with probability $1/64$ or $1/256$. Passive querying chooses a random stage of a current-student rollout. A state-stratified comparator spends half its queries on sensitive prompts at the final gate, using the public flag and trailing-action state; it does not inspect the teacher label when choosing a query. Rejected prompt draws and unsuccessful gate-reaching rollouts are counted. Cumulative label budgets are 32, 128, and 512 full four-action conditionals. The ordinary transformer receives the raw tokens. A typed-gate comparator additionally preserves the base outside the final gate and uses separate learned offsets for the two prompt flags, with residual width 0.1.

We evaluate each trained policy on 1,024 independent rollouts of each prompt type. Table 1 reports risk *conditional on a sensitive prompt*; unconditional failure is this number times the prompt frequency. This avoids making a system look safe merely because hazards are rare. Teacher mismatch and execution rate on ordinary prompts are also reported. A transformer that reduces risk by changing every sensitive action, rather than only the intended final decision, need not preserve teacher behavior or utility. The gated comparator makes that distinction testable by preserving the other decisions exactly.

Labels, behavior, and sampling cost. At frequency $1/256$ and 128 queries, all three passive ordinary-transformer runs miss the critical exception. Their mean conditional risk is 97.92%. The targeted transformer receives 64 exception labels on average and reduces risk to 0.07%, but executes on only 18.83% of sensitive-prompt steps. Its teacher KL is 0.83 nats per visited decision. By contrast, the targeted typed-gate model executes on 52.62% of steps, matches the conditional targets to numerical precision, and has risk 8.46% against the teacher's 8.91%. Its parameter certificate is 12.92%. Lower violation risk alone would give the wrong ranking for a distillation objective that also preserves the teacher's useful behavior.

Both targeted learners preserve ordinary-prompt behavior in this example: their mean execution fractions are approximately 0.65, the base value. Their difference is within sensitive prompts, where the unconstrained transformer changes decisions that the teacher would leave untouched. Exact gating prevents that particular overgeneralization. Execution frequency is a transparent behavior diagnostic, not a universal definition of task utility; the separate placement study uses explicit completion labels.

The mean prompt-draw costs for the targeted ordinary and typed learners are 31,400 and 15,870, respectively, at the same 128-label budget. The ordinary learner's low execution rate makes the critical state harder to reach for later targeted queries. Thus query design and learned behavior affect subsequent collection cost. At 512 labels, every method is evaluated again rather than estimated from

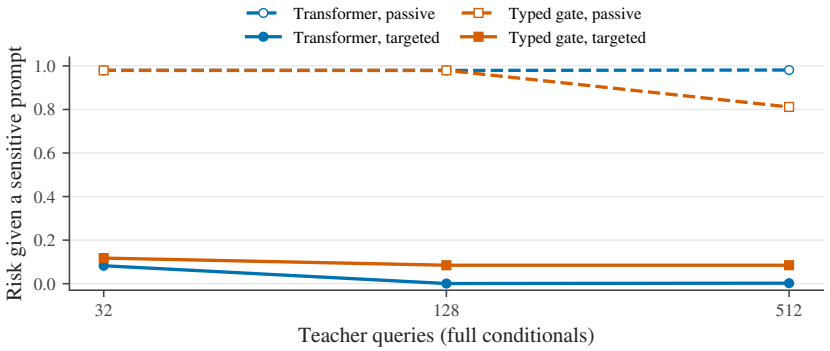

Figure 2: Learned query-budget curves at sensitive-prompt frequency $1/256$. Means are over three seeds, with 1,024 sensitive-prompt evaluation episodes per checkpoint. Targeting supplies the rare labels; gating preserves decisions outside the critical state. The ordinary targeted model's lower risk comes with substantial teacher mismatch, as discussed in the text.

Table 2: Coarse-state residual sweep, $H = 32$, 512 queries. Risks and bounds are percentages. Teacher KL is reported in 10^{-3} nats (values below 0.01 are shown as < 0.01) and averages the two mode-specific on-policy conditional KL means. Exact tying ($\gamma = 0$) cannot represent mode-specific targets; a residual improves fit, while its parameter box becomes looser. The refined gate uses the observable mode in its offset.

γ / gate	Teacher KL (10^{-3} nats)	Risk mode 0	Risk mode 1	Bound mode 1
0	14.74	39.19	39.19	39.60
0.25	3.18	64.10	33.63	65.41
0.5	0.04	82.00	26.24	82.96
1	0.04	80.83	26.82	94.99
2	0.09	80.63	27.08	99.97
4	0.52	77.38	26.86	100.00
Refined (0.1)	< 0.01	81.35	27.08	38.00

the 128-label result. Figure 2 shows the mean budget curves at frequency $1/256$. The three-seed study establishes the mechanism on these tasks, not a precise population rate across arbitrary prompts.

Targeted access is not free. At frequency $1/256$, constructing the stratified labels requires orders of magnitude more unlabeled prompt draws than the teacher-query count. The protocol supplies a known relevant flag; discovering such a flag is a different learning problem. The independent rare-cell experiment in Appendix C separately tests the missed-cell formula, while the transformer experiment tests what learning does after a useful label is, or is not, observed.

5.2 A COARSE GATE AND A USEFUL RESIDUAL

We next use $H = 32$ and equally frequent prompt modes. Both modes regard three consecutive executions as a violation, but the teacher's critical execution probability is 0.25 in one mode and 0.05 in the other. Outside the critical state it remains the base. The coarse gate records trailing-execution count and ignores mode; a tied conditional therefore cannot match both teacher distributions. The residual transformer sees the full prompt, including mode.

We sweep $\gamma_1 \in \{0, 0.25, 0.5, 1, 2, 4\}$ at 512 stratified teacher queries. The shared intercept is initialized from the queried gate labels before residual optimization, preventing a constant-logit saturation basin from being mistaken for a representational limit. A refined typed gate is included with the same label budget. Each coarse model has a valid certificate derived from its parameters, regardless of optimization success.

Teacher accuracy and certificate tightness. The teacher's exact risks are 82.20% and 27.30% in the two modes. Exact tying compromises between their critical distributions: both learned risks

Table 3: Risk (%) at exactly 95% expected completion on eight held-out tasks, averaged over two bases per task. Utility-aware rows use the single multiplier $\lambda = 1$. Each deployed mixture uses one selected set plus common base/abort options, so its supervised-context union is at most B. All methods have full teacher/task access.

Fixed single-set rule	$B = 4$	$B = 8$	$B = 16$
Teacher-cost	28.51	17.35	3.62
Linear, risk only (0.01)	27.01	18.41	1.11
Linear anchored	20.57	6.49	1.76
Teacher unsigned + utility	24.37	11.54	1.85
Teacher signed + utility	19.31	4.57	1.82
Exact scalarized	18.54	4.57	1.77

become 39.19%, and the mean teacher KL is 1.47×10^{-2}. A residual width of 0.5 reduces mean KL below 10^{-4} and separates the risks to 82.00% and 26.24%. The residual is therefore carrying behavior that the coarse cell offset cannot express. Its universal bound for the safer mode is nevertheless 82.96%, because the coarse logit box permits the less-safe mode's behavior there.

Increasing the width to 1 retains close teacher imitation but raises that bound to 94.99%; at width 2 it is nearly 100%. This is the cost of uncertain within-cell dependence, not evidence that the trained model itself becomes equally unsafe. Adding the observed mode to the cell definition produces a 38.00% bound for the safer mode with residual width 0.1. Exact state refinement is therefore valuable information, whereas parameter sharing alone does not justify treating different contexts as behaviorally identical. A larger residual is useful when fidelity requires it; a tighter certificate then requires a better state description or an independent risk evaluation.

The required residual width can also be calculated. Two full-support categorical targets p^0, p^1 fit a common offset and closed per-logit boxes of width γ exactly when

$$\gamma \geq \frac{1}{4} \left\{ \max_a \log \frac{p_a^0}{p_a^1} - \min_a \log \frac{p_a^0}{p_a^1} \right\}. \tag{8}$$

For finite $\tanh$ outputs the nondegenerate boundary is an infimum; any strictly larger width suffices for the logit representation. The current teachers have threshold approximately 0.46, consistent with the fit transition between 0.25 and 0.5. This is a representation threshold, not a training-convergence guarantee. Appendix B gives the short necessity and sufficiency argument.

The important comparison is not whether a bigger residual always lowers risk. It is whether it improves teacher fit through differences within a cell, and what happens to the certified risk. Equation (7) widens as residual freedom increases. Thus a coarse representation can support good conditional imitation while yielding a weak universal certificate. Refining the trusted state with an observable feature reduces this gap, but spends structural information rather than teacher labels.

5.3 Fair placement at matched completion

The finite-tree control uses 12 depth-five tasks and two trained neural bases per task. All selectors receive the same complete teacher and task model. A selected set is deployed by copying its teacher conditionals exactly and keeping the base elsewhere. These are counterfactual conditional policies, not newly trained networks or query-efficient algorithms. Four tasks choose tuning parameters; the other eight are held out for reporting. A fixed multiplier $\lambda = 1$ is evaluated separately.

For each single set, its policy, the unmodified base, and a safe-abort policy form a three-member bank. An episode-level mixture attains exactly 95% expected completion. Its union of teacher-supervised contexts is at most B. The exact planner minimizes $R - \lambda U$; the linear anchored rule uses the base-occupancy derivative of this objective. Teacher-cost uses teacher visitation times unsigned failure sensitivity. We also test stronger teacher-based rules using utility and the signed teacher–base displacement:

$$S_h^{\mathrm{TS}} = d_p(h)(b_h - p_h)\{A_p^R(h) - \lambda A_p^U(h)\}. \tag{9}$$

Selecting the largest positive scores gives a particularly simple comparator.

On the eight held-out tasks, the fixed signed teacher rule improves over ordinary teacher-cost by 12.78 ± 4.88 percentage points at budget eight, where the interval is a paired 95% Student-t interval over task means. The fixed linear anchored rule improves by 10.87 ± 4.06 points. The risk-only radius-0.01 linear rule instead differs by $+1.06 \pm 6.58$ points. Thus substituting a base envelope while leaving the selection objective unchanged is not sufficient in this control. The utility term and signed change in the action distribution supply the useful information.

The signed utility-aware heuristic nearly matches exact planning at budget eight with the single fixed multiplier. Accounting for the *direction* of imitation and for completion is important. A multiplier selected on four validation tasks is not uniformly better than the fixed one, particularly at budget sixteen.

5.4 A CONCRETE INDEPENDENT-QUERY CHECK

The tied three-cell calculation in Section 4 is evaluated separately from adaptive transformer learning. For each of six query counts we generate 20,000 independent batches from the stated fixed cell distribution. At 22 labels, the sufficient bound gives at least 95% probability of observing all cells. Inclusion–exclusion gives the sharper exact coverage probability 98.85%; the observed fraction is 98.80%. At 16 labels the exact missing-cell probability is already 4.43%, illustrating the conservatism of the simple minimum-mass bound. These are directly computable guarantees for the stated class, not estimates of a supremum over a continuous neural hypothesis class.

The rare-cell extension independently varies prompt frequency and query count. At frequency $1/1024$, the sensitive critical cell has query mass $0.56/1024$. With 512 independent labels, its exact missed-cell probability is 75.57%, versus 75.54% observed across 20,000 batches. Even with 2,048 labels it remains 32.62%. A network may still guess the exception, or refuse broadly, but the labels alone cannot reliably distinguish two teachers that differ only there. This calculation explains the access limitation; the adaptive experiment in Figure 2 separately measures the consequences of fitting and generalization after query selection. We do not apply the independent-sample bound to those adaptive trajectories.

5.5 DIAGNOSTICS THAT CHANGE THE INTERPRETATION

The radius evaluation reuses 1,728 trained policies at five radii, giving 8,640 policy–radius conditions. At base radius 0.01, the raw certificate exceeds observed risk in only 79.31% of conditions. Adding *both* selected and unselected residual corrections restores a valid finite-tree bound in every checked case. Mean within-task/init/budget rank correlation is 0.74 for nonconstant raw certificates and 0.79 for corrected bounds; 25% of raw-certificate groups are constant because the group has full supervision. These observations rule out reading the unverified radius as a deployment guarantee.

We also examine the local-selection theorem's sufficient conditions, rather than only its ranking accuracy. Its common-radius placement-regret bound is $2H(H-1)\varepsilon^2$. In the 1,200 KL-ranking conditions, the sufficient identification threshold has median 1.02×10^{-5} and maximum 3.29×10^{-4}; the smallest tested radius is 10^{-3}. Therefore *none* is inside that sufficient regime. The local result serves as a supporting approximation in Appendix D, not a quantitative explanation of the far-teacher experiments. The biased-routing theorem instead holds at the finite radii used in its structural checks.

6 RELATED WORK

Distillation transfers conditional or sequence distributions (Hinton et al., 2015; Kim & Rush, 2016; Ranzato et al., 2016). Generalized knowledge distillation (GKD), MiniLLM, and divergence-based sequence distillation separate rollout distributions from fitting objectives (Agarwal et al., 2024; Gu et al., 2024; Wen et al., 2023). On-policy self-distillation for safety alignment (OPSA) uses privileged-context teacher supervision on student trajectories; SafeSteer restricts the penalized vocabulary coordinates; soft-prompt safety distillation uses distributional objectives (Fu et al., 2026; Li et al., 2026a; Alfarra et al., 2026). Our local full-conditionals concern *which histories* receive labels, while the gate controls where an update may generalize. None of those losses alone implies the proposed pointwise envelope.

Compounding-error analyses and dataset aggregation (DAgger) distinguish teacher-visited from learner-visited contexts (Ross & Bagnell, 2010; Ross et al., 2011; Rajaraman et al., 2020). Cost-sensitive imitation and safety-guided collection motivate consequence-weighted querying (Ross & Bagnell, 2014; Ciftci et al., 2024); SafeDAgger and random network distillation DAgger (RND-DAgger) reduce teacher access through safety decisions or novelty (Zhang & Cho, 2017; Biré et al., 2025). Our metadata-stratified comparator is deliberately simple and assumes the relevant flag is available. Its purpose is to expose the rare-label bottleneck, not to replace active imitation algorithms.

Baseline bootstrapping and its factored and soft variants preserve a reference policy where change is uncertain (Laroche et al., 2019; Simão & Spaan, 2019; Nadjahi et al., 2020). Our gate uses the same general principle but certifies a declared region directly from the architecture, allowing an unrestricted cell offset where adaptation is admitted. State equivalence supplies the finite-state verifier (Givan et al., 2003); arbitrary parameter ties are coupled uncertainty and need not admit rectangular evaluation (Wiesemann et al., 2013; Li et al., 2026b). Machine teaching and alignment verification also distinguish behavior identification from evidence sufficient for a property (Brown & Niekum, 2019; Brown et al., 2021). The local expansion, occupancy constraints, and evaluations use standard policy-comparison, constrained-control, and statistical tools (Schulman et al., 2015; Altman, 1999; Achiam et al., 2017; Clopper & Pearson, 1934). The main structural result here is the exact extremal-router reduction and its coverage consequence; the architecture and experiments address whether its assumptions can hold in practice.

7 Discussion

A useful certificate requires an established envelope, not merely an appealing robust objective. With a trusted observable gate, large teacher-directed changes can coexist with strict preservation elsewhere. With a coarse gate, useful residual variation can loosen the certificate. With rare exceptions, even a compact student needs evidence at the relevant cells. For known finite tasks, a signed utility-aware local score is a strong practical baseline; the fair controls give no reason to insist on an exact worst-case planner.

Practical steps. First specify the failure predicate and which input features can be trusted. When a tool protocol supplies a valid adaptation gate, preserve the base outside it and fit separate offsets or a bounded residual inside it. This establishes an envelope without enumerating every token history. Second, estimate whether the available query process reaches each safety-relevant cell: a compact representation does not eliminate rare-label access. Metadata-based targeting is worthwhile only with its collection cost included. Third, compare selection rules at the same completion and the same union of supervised contexts. Finally, use the parameter certificate only at its declared state resolution; if it is too loose, refine that resolution, reduce residual freedom, or report a direct evaluation rather than silently assuming a smaller drift radius.

The theory concerns explicit failure predicates and specified behavioral classes. The biased-routing result exploits repeated hazard structure and monotonicity; its entropy exponent is a context-certificate property, not a neural query lower bound. The architectural guarantee trusts the parser, gate, transitions, and learned parameters. Our transformers are small, trained from scratch on synthetic commands, and not pretrained language models. Stratified access uses known metadata and incurs reported unlabeled-sampling cost. Statistical coverage claims use their stated independent-sample setting; adaptive learning curves and counterfactual placement controls provide separate evidence. These boundaries identify what must be supplied before a selective-distillation claim can be used operationally.

AI USE STATEMENT

In this work, we used generative AI tools to draft sections of the paper and polish writing. Additionally, we used generative AI tools for editing code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31, 2017. URL https://proceedings.mlr.press/v70/achiam17a.html.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13649.

Motasem Alfarra, Cristina Pinneri, Dana Kianfar, Mohammed Almousa, and Christos Louizos. Distilling safe LLM systems via soft prompts for on device settings. *arXiv preprint arXiv:2606.09388*, 2026. URL https://arxiv.org/abs/2606.09388. Accepted to the 42nd Conference on Uncertainty in Artificial Intelligence, 2026.

Eitan Altman. *Constrained Markov Decision Processes*. Chapman & Hall/CRC, 1999. ISBN 9780849303821. URL https://www.routledge.com/Constrained-Markov-Decision-Processes/Altman/p/book/9780849303821.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://research.google/pubs/conformal-risk-control/.

Emilien Biré, Anthony Kobanda, Ludovic Denoyer, and Rémy Portelas. Efficient active imitation learning with random network distillation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3f9f3159318fd946e1ddf510e059a3f6-Abstract-Conference.html.

Daniel S. Brown and Scott Niekum. Machine teaching for inverse reinforcement learning: Algorithms and applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7749–7758, 2019. doi: 10.1609/aaai.v33i01.33017749.

Daniel S. Brown, Jordan Schneider, Anca Dragan, and Scott Niekum. Value alignment verification. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1105–1115, 2021. URL https://proceedings.mlr.press/v139/brown21a.html.

Yusuf Umut Ciftci, Darren Chiu, Zeyuan Feng, Gaurav S. Sukhatme, and Somil Bansal. SAFE-GIL: SAFEty guided imitation learning for robotic systems. *arXiv preprint arXiv:2404.05249*, 2024. URL https://arxiv.org/abs/2404.05249.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2 edition, 2006. doi: 10.1002/047174882X.

Awi Federgruen and Henri Groenevelt. The greedy procedure for resource allocation problems: Necessary and sufficient conditions for optimality. *Operations Research*, 34(6):909–918, 1986. doi: 10.1287/opre.34.6.909.

Yu Fu, Longxuan Yu, Haz Sameen Shahgir, Zhipeng Wei, Hui Liu, N. Benjamin Erichson, and Yue Dong. Reducing the safety tax in LLM safety alignment with on-policy self-distillation. *arXiv preprint arXiv:2605.15239*, 2026. URL https://arxiv.org/abs/2605.15239.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003. doi: 10.1016/S0004-3702(02)00376-4.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005. doi: 10.1287/moor.1040.0129.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1139.

Romain Laroche, Paul Trichelair, and Rémi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3652–3661, 2019. URL https://proceedings.mlr.press/v97/laroche19a.html.

Hao Li, Jingkun An, Zijun Song, Pengyu Zhu, Rui Li, Hao Wang, Wendi Feng, Yesheng Liu, Lijun Li, Jin-Ge Yao, and Lei Sha. SafeSteer: Localized on-policy distillation for efficient safety alignment. *arXiv preprint arXiv:2606.02530*, 2026a. URL https://arxiv.org/abs/2606.02530. Version 2, August 2026; accepted to the 2026 Conference on Empirical Methods in Natural Language Processing.

Mengmeng Li, Daniel Kuhn, and Tobias Sutter. Policy gradient algorithms for robust MDPs with nonrectangular uncertainty sets. *SIAM Journal on Optimization*, 36(1):120–151, 2026b. doi: 10.1137/24M1631250.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

Kimia Nadjahi, Romain Laroche, and Rémi Tachet des Combes. Safe policy improvement with soft baseline bootstrapping. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2019)*, volume 11908 of *Lecture Notes in Computer Science*, pp. 53–68. Springer, 2020. doi: 10.1007/978-3-030-46133-1_4.

Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. doi: 10.1287/opre.1050.0216.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hTEGyKf0dZ.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6Mxhg9PtDE.

Nived Rajaraman, Lin F. Yang, Jiantao Jiao, and Kannan Ramchandran. Toward the fundamental limits of imitation learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2914–2924, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1e7875cf32d306989d80c14308f3a099-Abstract.html.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.06732.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 661–668, 2010. URL https://proceedings.mlr.press/v9/ross10a.html.

Stéphane Ross and J. Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning. *arXiv preprint arXiv:1406.5979*, 2014. URL https://arxiv.org/abs/1406.5979.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

Thiago D. Simão and Matthijs T. J. Spaan. Safe policy improvement with baseline bootstrapping in factored environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4967–4974, 2019. doi: 10.1609/aaai.v33i01.33014967.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. f-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10817–10834, 2023. doi: 10.18653/v1/2023.acl-long.605.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013. doi: 10.1287/moor.1120.0566.

Jiakai Zhang and Kyunghyun Cho. Query-efficient imitation learning for end-to-end simulated driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. doi: 10.1609/aaai.v31i1.10857.

A PROOFS OF THE ROUTING RESULTS

A.1 REMOVING ROUTER SUPERVISION

Every routing node has two isomorphic remaining subproblems. Suppose their selected sets produce robust conditional risks $x \le y$. A selected routing interval contains $1/2$, so its backup is at least $(x+y)/2$, whatever the assignment of its children. Delete that routing constraint and, when necessary, exchange the entire selected patterns in the two child subtrees so that the child of action 1 has risk x. All leaf labels, local interval choices, and remaining horizons match under this exchange. The unselected backup is

$$\max_{q\in[\ell,r]} (1-q)y + qx = (1-\ell)y + \ell x \le (x+y)/2.$$

The operation uses one fewer constraint. Applying it bottom up eliminates every selected router without increasing the root risk. At-most budgets are important: deletion need not be replaced by another selected context. Constraints on off-hazard branches cannot affect a failure, so they may also be removed.

This argument permits arbitrary selected-router intervals containing $1/2$, including a KL ball centered at $1/2$. It does not require the selected interval to be nested in $[\ell, r]$. It relies on repeated identical suffix problems and the one-sided bias $\ell > 1/2$; it is not a general claim that supervising routing decisions is useless.

A.2 MATCHING THE FIXED-ROUTING LOWER BOUND

After deleting router constraints, let $n_x \in \{0, \dots, k\}$ be the number of selected all-one hazard gates in copy x. Their conditional robust risk is $L(x) = u^{n_x} v^{k-n_x} = v^k a^{n_x}$, $a = u/v$. A feasible stationary adversary uses probability ℓ at every router, giving

$$R_\ell(n) = \sum_x \ell^{|x|_1} (1-\ell)^{m-|x|_1} v^k a^{n_x}.$$

Thus the minimum of this expression is a lower bound on the optimal robust certificate.

For fixed routing weights, increasing n_x from j to $j+1$ yields gain $G(x,j) = w_x v^k (1-a) a^j$. Gains within a copy strictly decrease. Selecting the largest B gains across copies automatically satisfies each chain's prerequisite: if gain j is included, all gains $0, \dots, j-1$ are larger and can also be included. At ties one may include a prefix within a copy. This proves the separable allocation formula. The same argument works for fixed arbitrary routing weights and copy-specific decreasing gain sequences; it is the standard resource-allocation component (Federgruen & Groenevelt, 1986).

The special robust claim needs more. If x' is obtained by changing one zero in x to one, then $w_{x'} > w_x$. For any stage j, $G(x', j) > G(x, j)$. In a top-gain allocation, including stage j in copy x therefore implies including stage j in copy x'. Consequently $n_{x'} \ge n_x$, including when some other gains tie across different stages. Hence $L(x') \le L(x)$.

Any coordinatewise nonincreasing terminal function on a binary cube has its maximal expectation over adaptive conditional probabilities in $[\ell, r]$ at the stationary lower endpoint ℓ. To prove this, compare the two functions on the last coordinate: $L(z,1) \le L(z,0)$, so the last-step maximizing probability is ℓ. Their expectation remains coordinatewise nonincreasing in the preceding coordinates, allowing induction. Equivalently, couple every conditional Bernoulli action with a Bernoulli-ℓ action using the same independent uniforms; each actual action is at least the baseline action, and monotonicity bounds the terminal loss. The induction argument also directly accommodates history-dependent choices. This proves that the top-gain allocation attains its fixed-routing lower bound against every feasible adaptive router and completes Theorem 1.

A.3 COMPRESSED COMPUTATION AND NUMERICAL CHECKS

At Hamming weight t, each routing string has weight $\ell^t (1-\ell)^{m-t}$, and there are $\binom{m}{t}$ such strings. For every $t \in \{0, \dots, m\}$ and $j < k$, form one gain value with multiplicity $\binom{m}{t}$. Sort the $k(m+1)$ types and accumulate gains until the desired reduction $v^k - r$ is reached. If the target falls within

a type, take the ceiling of the remaining reduction divided by its per-context gain. The number of arithmetic operations is $O(k(m+1)\log(k(m+1)))$; integer counts require bit lengths growing with m. This is a compressed target-budget computation, not a polynomial algorithm in the bit length of all possible real-valued inputs. The numerical computation handles interior targets and the all-supervised endpoint separately.

For independent checks we use the generic repeated-tree recurrence. Set $F_0(B) = u^{\min(B,k)}v^{k-\min(B,k)}$. For $d \geq 0$, write $I^0 = [\ell, r]$ and let I^1 be the selected-router interval. Then

$$F_{d+1}(B) = \min_{\substack{s\in\{0,1\},\ s\leq B \\ i+j\leq B-s}} \Phi_{I^s}\big(F_d(i), F_d(j)\big). \tag{10}$$

This recurrence does not use the biased-routing theorem, permits supervising a router, and takes $O(mB^2)$ arithmetic work when all budgets up to B are computed. It agrees with full explicit-tree placement on small cases and with Theorem 1 across 120 random models, six targets each. Each random model uses $k = 3$, $m \in \{1, \ldots, 6\}$, $\ell \in (0.51, 0.9)$, $v \in (\ell, 0.99)$, and $u \in (0.5, \min(v, 0.8))$. The two independent full frontiers agree within 10^{-14}. An additional 189 compressed cases reach $m = 256$ and vary base, base radius, and target mass.

When uniform routing is feasible, the exact recurrence still applies but the entropy reduction is not asserted. In the symmetric interval family $[1-v, v]$ and $[1-u, u]$, a uniform adversary gives the familiar integer-target budgets j for an early hazard and $j2^m$ for a late hazard. The latter follows because each of the 2^m copies must average at least j constraints at target $v^{k-j}u^j$; assigning j in every copy is sufficient. Relative to explicit tree size, this is a constant-fraction coverage requirement, not a neural sample lower bound. Between these targets, averaging two adjacent copy budgets need not attain the robust value: an adversary can favor the less-protected copy. Equation (10) resolves every intermediate target without that assumption.

A.4 Coverage mass and the entropy boundary

Let $N_w(\beta)$ be the smallest size of a set with stationary routing mass at least β. If a placement touches only copies in S, its reduction under stationary routing is at most $(v^k - u^k)\sum_{x\in S} w_x$. Reaching target $v^k - \beta(v^k - u^k)$ therefore requires at least $N_w(\beta)$ touched copies, and at least that many constraints. Conversely, fully supervising the $N_w(\beta)$ highest-weight copies gives the target using $kN_w(\beta)$ constraints. The selected set is an upper set under bitwise order: higher Hamming weight has strictly higher routing weight, and arbitrary ties inside a weight level preserve this property. Its residual loss is coordinatewise nonincreasing, so the same upper bound is valid for adaptive routing, not just the stationary lower bound.

For completeness, write $h = h_2(\ell)$. For every $\zeta > 0$, the Bernoulli typical set has probability tending to one and contains at most $2^{m(h+\zeta)}$ strings. It thus carries mass β for large m. In the other direction, outside a set of probability tending to zero, every individual string has probability at most $2^{-m(h-\zeta)}$. Any set with fewer than $2^{m(h-2\zeta)}$ strings has mass tending to zero. These two standard equipartition bounds (Cover & Thomas, 2006) give $\log_2 N_w(\beta)/m \to h$. Multiplication by fixed k does not change the exponent, proving Corollary 2. The endpoint $\beta = 1$ differs because every string has positive probability for $\ell \in (1/2, 1)$. To attain the minimum possible risk u^k, every copy must have all k constraints. At $\ell = 1$, only one routing string is reachable, so that endpoint statement would fail; this deterministic case is excluded from the theorem.

B Architectural envelopes and verification

B.1 KL control outside the gate

Fix a full-support base vector b and perturbation s with $\|s\|_\infty \leq \gamma_0$. Define $q_t(a) = b(a)e^{ts_a}/Z(t)$ and $A(t) = \log Z(t)$. Then

$$D_{\mathrm{KL}}(q_1\|b) = A'(1) - A(1), \qquad \frac{d}{dt}\{tA'(t) - A(t)\} = tA''(t) = t\,\mathrm{Var}_{q_t}(s).$$

The perturbation range is at most $2\gamma_0$, so its variance is at most γ_0^2. Integrating $t\gamma_0^2$ over $[0, 1]$ proves $D_{\mathrm{KL}}(q_1\|b) \leq \gamma_0^2/2$. This calculation holds pointwise for every network input and every parameter

setting. Zero base probabilities remain zero under finite logit tilts; the implementation uses a positive base. If only a single binary logit is perturbed by γ_0, its sufficient KL bound improves to $\gamma_0^2/8$ because the two-action perturbation range is only γ_0.

The gate-dependent construction enforces this property on the complement of C while using unrestricted offsets α_g inside C. Thus there is no conflict between exact base preservation outside and a distant teacher target inside. The construction does not decide which generalizations are justified. That choice is represented by the supplied gate and its features, which must be correct for the claimed protocol.

B.2 The width needed to fit two within-cell targets

Let $d_a = \log p_a^0 - \log p_a^1$ and $D = \max_a d_a - \min_a d_a$. If $p^i = \mathrm{softmax}(\alpha + s^i)$ with $|s_a^i| \le \gamma$, the log-ratio vector differs from $s^0 - s^1$ by only an action-independent normalization constant. Its range is at most 4γ, proving necessity of (8). Conversely set $c = (\max d + \min d)/2$, $\alpha = (\log p^0 + \log p^1)/2$, $s^0 = (d - c)/2$, and $s^1 = -s^0$. Each residual has maximum absolute value $D/4$, while $\mathrm{softmax}(\alpha + s^i) = p^i$. This proves sufficiency for the closed box. A finite tanh residual occupies the open box, so at nonzero D equality is approached as a limit and strictly larger widths permit finite preactivations. A network must still be able to distinguish the two contexts and learn those residuals.

When the nonexecution action proportions are the same, $D = |\operatorname{logit}(e_0) - \operatorname{logit}(e_1)|$. With execution targets 0.25 and 0.05, $D/4 \approx 0.46$. This diagnostic explains why width 0.25 cannot exactly fit the two distributions whereas width 0.5 can. Twenty random categorical-pair tests reconstruct the targets from the explicit optimal offsets and residuals. This elementary logit geometry is used as a mechanism check, not a general new result about neural expressivity.

B.3 Parameter intervals and a coarse state model

The coordinate probability $\mathrm{softmax}(z)_a$ increases with z_a and decreases with every z_j, $j \neq a$. Its extrema on $z_j \in [\alpha_j - \gamma_1, \alpha_j + \gamma_1]$ are therefore the two opposing corners in (7). Separate coordinate bounds need not describe a simplex exactly. In our grammar only execution versus nonexecution affects the next safety state, so the exact coordinate interval for execution is sufficient. For a general multi-action transition model, maximizing over the logit box or a simplex-compatible outer approximation is required; independently setting all action probabilities to their separate upper endpoints would be invalid.

The safety state is the number $s \in \{0, 1, 2\}$ of trailing executions, plus an absorbing failure flag. Nonexecution resets s to zero. Execution increments s or causes a first failure from state two. For a conditional probability interval $[l_s, u_s]$, starting at $V_0(s) = 0$, back up

$$V_{t+1}(s) = \max_{q \in [l_s, u_s]} (1 - q) V_t(0) + q\big(\mathbf{1}\{s = 2\} + \mathbf{1}\{s < 2\} V_t(s + 1)\big).$$

A fixed prompt mode can be included in the finite state even when the parameter cell g ignores it. Thus the state model can track the failure predicate correctly while the learned head has a coarser abstraction of teacher behavior. The bound remains valid because every network output belongs to its parameter interval. Time sharing and within-cell network constraints may make the bound conservative. Exact interval backups do not imply exact characterization of all networks in that box.

For the binary bounded-residual GRU comparison, the parameter interval is $[\sigma(\alpha_s - 0.1), \sigma(\alpha_s + 0.1)]$. The representative-expanded interval is $[q_s - 0.05, q_s + 0.05] \cap [0, 1]$. The former is contained in the latter because the sigmoid is $1/4$-Lipschitz and the representative belongs to the parameter interval. Both certificates are evaluated on the same 64 trained bounded-residual checkpoints; containment and the resulting certificate improvement are verified without retraining. For horizon 16, teacher critical probability 0.1, and 128 queries, means over four seeds are 36.52% and 43.34%, respectively. Neither parameter-only calculation requires running the recurrent residual on every history.

B.4 Aliasing and what cannot be inferred from a small loss

Let a deployment representation satisfy $\phi(h) = \phi(h')$, and restrict a policy to depend only on ϕ. If $\mathrm{TV}(q_h, p_h) \le e$ and $\mathrm{TV}(q_{h'}, b_{h'}) \le \kappa$, then $q_h = q_{h'}$ implies $\mathrm{TV}(p_h, b_{h'}) \le e + \kappa$. Whenever the

Table 4: Independent query batches on the three-cell class. Entries are percentages; positive values below 0.01% are shown as < 0.01. Each observed fraction uses 20,000 new batches. This is a direct test of the independent-sample coverage statement.

Queries	Exact miss (%)	Union bound (%)	Observed miss (%)
4	68.08	100.00	68.24
8	27.71	65.61	28.31
16	4.43	14.35	4.33
22	1.15	4.59	1.21
32	0.13	0.69	0.11
64	< 0.01	< 0.01	0.00

left side is larger, no learner in this class can simultaneously satisfy both requirements. The same conclusion follows from KL tolerances through Pinsker's inequality. This is an elementary but useful feasibility test before imposing a coarse gate.

It does not say a full-history transformer cannot learn a distinction. Rather, a deterministic guarantee for unobserved histories needs either a verified structural restriction, an independently justified generalization result, or direct verification. A finite collection of empirical losses alone leaves an unrestricted function free to change on an unseen input. In practice, tool identifiers and syntactic protocol states can be checked by a small parser. Learned dialogue-state trackers, ambiguous natural-language permission statements, and out-of-protocol inputs require a separate error model or a fallback. Those tasks are outside the present verifier.

C Query coverage: concrete bounds and their limits

C.1 Independent full-conditional queries on tied cells

Suppose both teacher and student are exactly tied on K known cells. Query locations are independent draws from a fixed cell distribution π, and the teacher reveals the full conditional vector. Let E be the event that every cell is hit. Then

$$\Pr(E^c) \le \sum_{g=1}^{K} (1-\pi_g)^n \le K \exp(-n\pi_{\min}).$$

On E, matching each observed vector establishes teacher equality on all cells. This statement holds for arbitrary continuous categorical vectors. With fitting total-variation error at most e in every cell, maximal coupling bounds the trajectory-risk difference by $1-(1-e)^H \le He$; the fitting condition must be checked separately. Approximate teacher sharing would add an additional within-cell discrepancy and is not inferred from one queried representative.

In the binary base process with execution probability 0.8, use the nonabsorbing trailing-execution state during query collection. From state probabilities d_t, the next probabilities are

$$d_{t+1} = (0.2 \sum_s d_t(s),\ 0.8 d_t(0),\ 0.8[d_t(1)+d_t(2)]), \quad d_0 = (1,0,0).$$

Sampling a stage uniformly from 0 to 15 gives average cell masses $(0.25, 0.19, 0.56)$. The exponential union bound gives $n \ge \lceil \log(3/0.05)/0.19 \rceil = 22$. We also compute the exact miss probability by inclusion–exclusion,

$$\Pr(E^c) = \sum_{\varnothing \ne S \subseteq [K]} (-1)^{|S|+1} \left(1 - \sum_{g \in S} \pi_g\right)^n,$$

and compare it with 20,000 independent multinomial query batches at each tested sample size. These draws satisfy the theorem's assumptions, unlike an adaptive training history.

C.2 A RARE EXCEPTION CANNOT BE IDENTIFIED BEFORE IT IS OBSERVED

Let two teacher hypotheses agree except on a cell of query probability η_*. Conditional on missing that cell, the full labeled query transcripts are identical. Under a uniform prior on the hypotheses, any identification procedure has error at least $1/2$ on this event. Its minimax error is therefore at least $\frac{1}{2}(1-\eta_*)^n$. To make this at most $\beta < 1/2$ requires

$$n \geq \frac{\log(1/(2\beta))}{-\log(1-\eta_*)}.$$

The exact inverse-mass expression, not an asymptotic class-size argument, is the relevant obstacle. With the preceding independent query distribution and a rare prompt flag of mass η, the critical cell has mass 0.56η. We test this formula for $\eta \in \{1/64, 1/256, 1/1024\}$, $n \in \{32, 128, 512, 2048\}$, and 20,000 independent batches each.

This is an identification lower bound. If failure labels are already known and the task permits refusing everything, safety alone needs no teacher label. Its relevance to distillation is preserving the teacher's context-dependent behavior and usefulness rather than applying a universal refusal rule.

C.3 A FINITE-CLASS RISK BOUND AS AN ILLUSTRATIVE ALTERNATIVE

For a fixed finite class $\mathcal{Q}$ with N policies and a fixed history distribution ν, define teacher advantage $A_p(h) = V_p(h1) - V_p(h0)$ and

$$C_\nu = \sup_{q \in \mathcal{Q}, h} \frac{d_q(h)|A_p(h)|}{\nu(h)}.$$

The teacher value telescope gives $R(q) - R(p) \leq C_\nu \mathbb{E}_\nu |q_h - p_h|$. For independent queries, one-sided Hoeffding and a union bound give, simultaneously over the predetermined class,

$$R(q) \leq R(p) + C_\nu \left[\frac{1}{n} \sum_{i=1}^{n} |q_{h_i} - p_{h_i}| + \sqrt{\frac{\log(N/\beta)}{2n}} \right].$$

Exact label consistency replaces the bracket by $\log(N/\beta)/n$ through a missed-disagreement-set argument. This is a direct use of standard concentration (Hoeffding, 1963). The coefficient can be infinite or very large for rich classes, so we do not apply the formula to continuous adaptive transformers. The cell-coverage calculation above is the concrete statistical bridge used in a studied restricted class. Sequential or phase-wise analysis would be needed to use adaptive queries in a formal learning guarantee (Howard et al., 2021).

D ENVELOPE DIAGNOSTICS AND THE LOCAL APPROXIMATION

D.1 THE QUANTITATIVE REGIME

Let $A_b(h) = V_b(h1) - V_b(h0)$ and assume every point in either local interval lies within probability distance ϵ_t of b_h at depth t. Define

$$C(M) = R(b) + \sum_h d_b(h) \max_{q \in I_h^{\mathbf{1}\{h \in M\}}} (q - b_h) A_b(h), \qquad E_H = 2 \sum_{t=0}^{H-1} \epsilon_t \sum_{s<t} \epsilon_s.$$

Then $|\mathcal{R}(M) - C(M)| \leq E_H$ uniformly in M. The exact local-proxy minimizer selects the largest positive differences of its two local maxima, and its certificate regret is at most $2E_H$. For common radius ϵ, this is $2H(H-1)\epsilon^2$. Replacing the loss by $y - \lambda c$ multiplies the remainder by at most $1 + \lambda$.

The proof is the standard performance-difference identity $R(q) - R(b) = \sum_h d_q(h)(q_h - b_h)A_b(h)$, followed by maximal coupling. At depth t, the history distributions differ in ℓ_1 by at most $2\sum_{s<t} \epsilon_s$. Replacing d_q by d_b costs at most E_H. Taking a maximum and then comparing selected-set minimizers costs at most twice the same remainder. The argument is closely related to trust-region and cost-sensitive policy comparison (Schulman et al., 2015; Ross & Bagnell, 2014); uniformity follows from applying the same bound to the whole interval family. We use it as a supporting approximation and not as a new general performance-difference technique.

Table 5: All 8,640 policy–radius conditions, with within-group rank correlations. Excess is measured in percentage points (pp). Coverage is an empirical fraction over known policies, not a probabilistic guarantee inferred from the average. Corrected bounds include both selected and unselected violations.

Radius	Quantity	Rank corr.	Constant (%)	Above risk (%)	Excess (pp)
0	bound	0.77	0.00	100.00	4.82
0	certificate	0.73	25.00	68.58	-0.43
10^{-3}	bound	0.78	0.00	100.00	5.89
10^{-3}	certificate	0.74	25.00	74.77	0.87
0.01	bound	0.79	0.00	100.00	8.45
0.01	certificate	0.74	25.00	79.31	3.78
0.1	bound	0.77	0.00	100.00	17.09
0.1	certificate	0.68	25.00	82.53	13.27
1	bound	0.49	0.00	100.00	33.49
1	certificate	0.27	25.00	83.58	30.96

D.2 KL RANKING AND AN EXPLICIT SUFFICIENT THRESHOLD

Let $p_h = b_h + \epsilon m_h$, $b_h \in [a, 1-a]$, $|m_h| \le M$, and radii $\epsilon^2 r_h^i$ with $0 \le r_h^i \le R$. The leading selected gain is

$$W_h = d_b(h)\left[|A_b(h)|\sqrt{2b_h(1-b_h)}(\sqrt{r_h^0} - \sqrt{r_h^1}) - A_b(h)m_h\right].$$

Put $D = M + \sqrt{R/2}$, require $\epsilon D \le a/2$, and set

$$T = \frac{4}{a^2(1-a/2)^2}, \quad K_0 = \frac{TR}{24} + \frac{M\sqrt{R}}{\sqrt{a(1-a/2)}}, \quad K = HK_0 + H(H-1)D^2.$$

Taylor expansion of binary KL on $[a/2, 1-a/2]$ bounds either endpoint's remainder by $K_0\epsilon^2$. Specifically, for $x = q-p$ and $\delta = \epsilon^2 r$, the third-derivative bound gives $|x^2 - 2p(1-p)\delta| \le T|x|^3/12$. Dividing by $|x| + \sqrt{2p(1-p)\delta}$ and using Pinsker's inequality yields endpoint error at most $T\delta/24$. The movement of the center contributes the second term in K_0. Combining endpoint and visitation remainders gives a uniform certificate remainder $K\epsilon^2$.

If the unique top-score set has gap Δ to the next feasible set, its exact certificate is optimal whenever $\epsilon < \Delta/(2K)$ and the interior condition holds. For the 80 depth-five soft-risk trees used to evaluate the approximation, $a = 0.15$, $M = 0.1$, $R = 1$, so $K \approx 57.66$. Across the 240 task–budget score sets, the gap ranges from 1.34×10^{-5} to 0.04, with median 1.18×10^{-3}. Thresholds range from 1.16×10^{-7} to 3.29×10^{-4}, with median 1.02×10^{-5}. None of the 1,200 tested radii, whose minimum is 10^{-3}, satisfies the sufficient threshold. The 97.92% set agreement at 10^{-3} is therefore an empirical asymptotic trend, not a demonstrated finite-radius guarantee. Each condition is checked using its score gap, remainder constant, sufficient threshold, set agreement, and regret.

D.3 UNSELECTED DRIFT, RESIDUALS, AND CORRELATION

The frozen-policy evaluation reuses 1,728 fine-tuned GRUs from 12 depth-five tasks, with task seeds 1000–1011 and two initializations per task. Each initialization contributes 72 policies: five selective methods at three budgets and four rehearsal weights, plus three full-context methods at four rehearsal weights. Re-evaluating these policies at base radii $0, 10^{-3}, 0.01, 0.1, 1$ gives 8,640 conditions without retraining. The independent task count is 12; policies and radii within a task are repeated measurements. For every proposed envelope, selected and unselected positive residuals are included in (2). The raw certificate alone need not exceed risk when the student violates its assumptions. Table 5 reports both the failure of that inference and the restored corrected bound. Correlations are calculated within task, initialization, and budget across that group's policies; constant groups are omitted from the correlation mean and counted explicitly. Full-supervision groups account for the constant raw values.

A residual diagnostic assumes access to the model conditionals and certificate values. Without that access, direct failure counting is the primary fixed-policy evaluation. Empirical Bernstein and time-

uniform methods can be useful under their respective sampling conditions (Maurer & Pontil, 2009; Howard et al., 2021); none converts an invalid pointwise base envelope into a valid one. Uniform risk-control methods similarly require an explicit data and selection protocol (Angelopoulos et al., 2024). The current paper does not claim an evaluation-efficiency advantage.