SELECTION-SAFE CONFORMAL EARLY EXIT: DISTRIBUTION-FREE COVERAGE UNDER ADAPTIVE DEPTH

Anonymous authors
Paper under double-blind review

ABSTRACT

Calibrating a conformal prediction set at each early exit does not preserve coverage after a data-dependent exit is selected. We construct one terminal path score from prefix maxima, calibrate one split-conformal quantile, and return nested prefix sets that all contain the valid terminal set. The resulting guarantee is selection-safe without an α/L correction and holds simultaneously for every post-calibration label-free policy that preserves terminal-set containment. We further bound the efficiency cost of the path maximum by the empirical tail of the early-over-terminal score excess, linking set inflation to the mechanism diagnostic $J(x, y)$. At the smallest admissible excess threshold, the held-out plug-in right-hand side is 0.016–0.086 against measured mean positive terminal-set inflation 0.01–0.08; the cell-wise ratio is 1.07–1.73. Across eight matched encoder task–backbone cells, including an RTE split with $n_{\mathrm{cal}} = 498$ drawn from the training partition, pathwise selection attains macro coverage 0.901 versus 0.814 for naive layerwise selection, with average set size 1.41 and latency fraction 0.71. On CIFAR-100 with a 24-exit ViT-L/16 and $n_{\mathrm{cal}} = 500$, first-eligible stopping at budget $b = 4$ attains coverage 0.902, average set size 2.76, and latency fraction 0.57; a design-chosen restriction reduces the returned set to 2.52 while, necessarily, increasing latency to 0.61. The same terminal calibration supports deployment-time budget changes without refitting. A joint L–n sweep isolates candidate-count scaling. At $n = 100$, pathwise set size changes from 3.02 to 3.11 as L increases from 4 to 24, whereas Bonferroni becomes exactly degenerate at $L \geq 12$ and returns all 100 classes. Against MinSE valid selection, the paired set-size difference at $(n, L) = (100, 24)$ is 0.80 $[0.47, 1.13]$, so the long-path, limited-calibration advantage is assessed from paired replicates rather than overlapping marginal intervals. Independent-depth CIFAR-100 and ImageNet-1k replications preserve the qualitative ordering. Full-vocabulary decoder experiments remain a negative boundary, with average sets containing $7{,}162$–$9{,}620$ tokens and nearly all examples reaching full depth. The empirical picture is therefore specific: pathwise calibration is most useful when the number of candidate exits is large relative to calibration size, when deployment budgets change after calibration, and when shallow heads can return actionable sets.

1 INTRODUCTION

Modern transformers are deep, but many inputs do not require every layer. Early-exit networks attach heads to intermediate representations and stop when a prediction appears sufficiently reliable (Teerapittayanon et al., 2016; Xin et al., 2020; Liu et al., 2020; Zhou et al., 2020). Their practical value depends on the stopping rule. Confidence, entropy, and prediction stability can identify easy examples, but these signals alone do not control how often the returned answer is correct (Guo et al., 2017; Kubaty et al., 2026).

Conformal prediction appears to offer an immediate remedy. One can calibrate a prediction set at each layer and return the first set small enough to act on. The difficulty is that the selected layer depends on the test input and the observed sequence of sets. Fixed-layer marginal coverage does not generally transfer to this selected set. Indeed, L individually valid exits can be arranged so that

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.

an earliest-singleton rule has miscoverage $L\alpha$. This is a structured instance of the post-selection problem studied for conformal predictors (Yang & Kuchibhotla, 2025; Sarkar & Kuchibhotla, 2023; Hegazy et al., 2025).

We make the complete depth path one conformal object. Let $r_\ell(x, y)$ be a fixed score for candidate label y at exit ℓ, and define

$$A_\ell(x, y) = \max_{j \le \ell} r_j(x, y). \tag{1}$$

We calibrate one quantile using the terminal score A_L. The set $\mathcal{C}_\ell(x) = \{y : A_\ell(x, y) \le \widehat{q}\}$ decreases with depth. Consequently, any returned prefix set contains $\mathcal{C}_L(x)$, an ordinary split-conformal set. Adaptive stopping therefore cannot reduce its coverage.

This construction separates three issues that are often conflated. First, it protects true-label coverage of the returned set, not agreement with a full-depth prediction. Second, it avoids a layerwise union bound because only one terminal score is calibrated. Third, it does not make weak early representations informative. A poor shallow score can enlarge the common threshold and erase the computational benefit. The empirical question is therefore not whether the guarantee holds, but whether it remains useful at matched coverage and measured latency.

The closest early-exit work makes this distinction important. EENN-AVCS constructs nested sets using anytime-valid confidence sequences and studies the posterior-stability approximations required for predictive inference (Jazbec et al., 2024a). Fast-yet-Safe controls a chosen supervised or reference-based loss (Jazbec et al., 2024b), while CALM controls degradation or consistency in adaptive language modeling (Schuster et al., 2022). Our narrower contribution is an exact split-conformal construction for the set returned after adaptive depth selection. It exploits the ordered path through pointwise containment rather than a general stability correction.

We contribute a counterexample to naive layerwise reasoning, an exact pathwise construction with one conformal rank, and an efficiency theorem that connects the cost of the path maximum to the tail of the early-over-terminal score excess. The same containment event makes the stopping policy a free post-calibration design variable: cardinality, margin, randomization, and load-aware policies may change without refitting as long as the frozen terminal set remains contained. We test the two consequences that distinguish this construction from generic safe selection: scaling in the number of candidate exits and deployment-time policy changes. A 24-exit CIFAR-100 benchmark supplies the non-binary regime in which set budgets are operational, while the full-vocabulary decoder study remains the opposing boundary.

2 Related Work

Conformal prediction converts fixed model scores into finite-sample marginal prediction sets (Vovk et al., 2005; Shafer & Vovk, 2008; Papadopoulos et al., 2002; Lei et al., 2018). Modern introductions summarize the resulting finite-sample construction (Angelopoulos & Bates, 2023). Cross-conformal prediction and jackknife+ improve data reuse (Vovk, 2015; Barber et al., 2021), while nested formulations clarify the connection between scores and set families (Gupta et al., 2022). APS, RAPS, and least-ambiguous classifiers target smaller classification sets (Romano et al., 2020; Angelopoulos et al., 2021; Sadinle et al., 2019). Full feature-conditional validity is impossible without assumptions (Vovk, 2012); Mondrian, weighted, and beyond-exchangeability methods instead protect specified partitions or shift models (Vovk et al., 2003; Tibshirani et al., 2019; Barber et al., 2023).

Selecting among marginally valid sets can invalidate coverage. Sample splitting, simultaneous inference, and stability corrections address general collections of predictors (Yang & Kuchibhotla, 2025; Sarkar & Kuchibhotla, 2023; Hegazy et al., 2025). Our setting is more structured: depth is ordered, and every admissible returned set is constructed to contain one terminal conformal set. Conformal risk control, risk-controlling prediction sets, and learn-then-test control more general losses and support the optional point-commit branch (Angelopoulos et al., 2024; Bates et al., 2021; Angelopoulos et al., 2025).

Conditional computation and adaptive computation allocate different amounts of work across inputs (Bengio et al., 2015; Graves, 2016; Figurnov et al., 2017). Early-exit architectures instantiate this principle in vision and language (Huang et al., 2018; Scardapane et al., 2020; Han et al., 2022; Schwartz et al., 2020; Elbayad et al., 2020).

DeeBERT, FastBERT, and PABEE use confidence, distillation, or prediction patience (Xin et al., 2020; Liu et al., 2020; Zhou et al., 2020); CALM and consistent accelerated inference calibrate model agreement (Schuster et al., 2021; 2022). EENN-AVCS and Fast-yet-Safe are the closest uncertainty-aware methods (Jazbec et al., 2024a;b). PAC-Bayesian analysis addresses generalization of adaptive depth rather than post-selection coverage (Guo et al., 2026).

Selective classification controls when a point prediction is returned (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; 2019). Conformal methods have also been applied to text classification, multiple-choice answering, and open-ended generation (Maltoudoglou et al., 2022; Kumar et al., 2023; Quach et al., 2024; Campos et al., 2024).

Cascaded conformal inference reduces candidate-evaluation cost across models (Fisch et al., 2021); our resource axis is the depth of one model. Coverage may additionally fail when a model changes after calibration (Shihab et al., 2026), so all backbones, heads, scores, and schedules are frozen before final calibration.

Table 2 separates the objects protected by the closest methods. This distinction matters because agreement with the deepest head, supervised classification risk, and coverage of a returned set are different estimands. A comparison is matched only when the controlled target and the information available to the stopping rule agree.

3 Selection-Safe Conformal Early Exit

3.1 Setup and the selection failure

A trained network has exits $\ell \in \{1, \dots, L\}$ and label space $\mathcal{Y}$. A head at exit ℓ produces logits $f_\ell(x)$. The training set fits the backbone and heads. An optional design split $\mathcal{D}_{\mathrm{des}}$ chooses exit locations and score transformations. The final calibration set $\mathcal{D}_{\mathrm{cal}} = \{(X_i, Y_i)\}_{i=1}^n$ is exchangeable with a new test pair (X, Y) conditional on everything fixed before calibration. The default score is $a_\ell(x, y) = -\log p_\ell(y \mid x)$.

An increasing map T_ℓ may be learned on $\mathcal{D}_{\mathrm{des}}$, after which $r_\ell = T_\ell \circ a_\ell$ is frozen.

Fixed-layer validity is insufficient. Let $Y = 1$ almost surely, and let $U \sim \mathrm{Unif}[0, 1]$ be part of X. For disjoint intervals $I_1, \dots, I_L$ of length α, exit ℓ returns $\{0\}$ on I_ℓ and $\{0, 1\}$ otherwise. Every exit has coverage $1 - \alpha$, but the first singleton is wrong on $\cup_\ell I_\ell$, giving coverage $1 - L\alpha$. The rule observes only X and set sizes. Appendix A gives the complete argument.

Several natural repairs do not yield the same result. Sharing one calibration split across exits creates dependence but does not make a selected marginal guarantee simultaneous. Taking a running intersection of independently calibrated sets enforces nesting, yet its terminal miss event is the union of the layerwise miss events and can therefore accumulate error. Bonferroni calibration at level α/L is valid for every exit simultaneously, but it protects a stronger event and can produce unnecessarily large sets. Finally, selecting score transformations or an exit schedule on the final calibration labels and then reusing those labels for the conformal rank breaks the simple exchangeability argument. The construction below instead conformalizes the object that the whole path must preserve.

3.2 One conformal rank for the path

Define the prefix maximum in Equation (1) and the terminal calibration scores $S_i = A_L(X_i, Y_i)$. With

$$k = \lceil (n + 1)(1 - \alpha) \rceil, \qquad \widehat{q} = S_{(k)}, \tag{2}$$

where $S_{(k)}$ is the kth calibration order statistic and $\widehat{q} = +\infty$ when $k = n + 1$, return

$$\mathcal{C}_\ell(x) = \{y \in \mathcal{Y} : A_\ell(x, y) \le \widehat{q}\}. \tag{3}$$

Because A_ℓ is nondecreasing in ℓ, $\mathcal{C}_1(x) \supseteq \cdots \supseteq \mathcal{C}_L(x)$. We use the first nonempty set of size at most a budget b, with terminal fallback; any other label-free rule is allowed if its returned set contains $\mathcal{C}_L(x)$. Table 4 summarizes the data-use order, and Appendix B gives pseudocode and the empty-set convention.

Theorem 1 (Selection-safe marginal coverage)**.** *Condition on the trained model, $\mathcal{D}_{\mathrm{des}}$, the transformations, and the exit schedule. If $\mathcal{D}_{\mathrm{cal}}$ and (X, Y) are exchangeable, then any label-free stopping rule whose returned set contains $\mathcal{C}_L(X)$ satisfies*

$$\Pr\{Y \in \mathcal{C}_{\widehat{\ell}(X)}(X)\} \geq 1 - \alpha.$$

The terminal scores and the test score are exchangeable, so the split-conformal rank argument gives $\Pr\{Y \in \mathcal{C}_L(X)\} \geq 1 - \alpha$. Pointwise nesting gives $\mathcal{C}_{\widehat{\ell}(X)}(X) \supseteq \mathcal{C}_L(X)$. This containment is the entire post-selection step; no independence across exits or layerwise union bound is used. Appendix C provides the formal proof.

Remark 1 (Policy-uniformity and the minimal envelope)**.** *The proof of Theorem 1 contains two immediate consequences that are useful operationally but require no separate theorem. First, on the event $Y \in \mathcal{C}_L(X)$, every label-free policy that returns a superset of $\mathcal{C}_L(X)$ covers the true label simultaneously. Thus, after $\widehat{q}$ is fixed, the system may change its cardinality budget or use an input-dependent, randomized, or load-dependent stopping rule without a new multiplicity correction, provided the scores, transformations, model, and exit schedule remain frozen. Second, for each fixed (x, y), $A_\ell(x, y) = \max_{j \leq \ell} r_j(x, y)$ is the pointwise minimal nondecreasing majorant of the exit scores: any nondecreasing sequence B_ℓ satisfying $B_\ell \geq r_\ell$ must obey $B_\ell \geq A_\ell$. This is a structural statement about irreversible candidate elimination, not a claim of global efficiency optimality after recalibration.*

Theorem 2 (Finite-sample path-maximum efficiency bound)**.** *For calibration example i, let*

$$B_i = r_L(X_i, Y_i), \qquad D_i = \left[\max_{j<L} r_j(X_i, Y_i) - r_L(X_i, Y_i)\right]_+,$$

so that the terminal path score is $A_i = B_i + D_i$. Let $B_{(k)}$ and $A_{(k)}$ be the conformal order statistics at the same rank k, and for $t \geq 0$ define $m_t = \sum_{i=1}^{n} \mathbf{1}\{D_i > t\}$. Whenever $k + m_t \leq n$,

$$0 \leq A_{(k)} - B_{(k)} \leq t + B_{(k+m_t)} - B_{(k)}. \tag{4}$$

Let $\mathcal{C}_L^{\mathrm{path}}$ use threshold $A_{(k)}$ with score A_L, and let $\mathcal{C}_L^{\mathrm{last}}$ use threshold $B_{(k)}$ with score r_L. Conditional on the calibration sample,

$$\left(|\mathcal{C}_L^{\mathrm{path}}(X)| - |\mathcal{C}_L^{\mathrm{last}}(X)|\right)_+ \leq \sum_{y \in \mathcal{Y}} \mathbf{1}\left\{B_{(k)} < r_L(X, y) \leq B_{(k)} + w_t\right\}, \tag{5}$$

where $w_t = t + B_{(k+m_t)} - B_{(k)}$. If the candidatewise CDF of $r_L(X, y)$ is M_y-Lipschitz on this interval, then

$$\mathbb{E}\left[\left(|\mathcal{C}_L^{\mathrm{path}}| - |\mathcal{C}_L^{\mathrm{last}}|\right)_+ \mid \mathcal{D}_{\mathrm{cal}}\right] \leq w_t \sum_{y \in \mathcal{Y}} M_y.$$

Moreover, if $p_t = \Pr\{D(X, Y) > t\}$, then with probability at least $1 - \delta$ over an i.i.d. calibration draw,

$$m_t \leq n p_t + \sqrt{\frac{n \log(1/\delta)}{2}}.$$

At $t = 0$, $p_0 \leq \Pr\{J(X, Y) < L\}$ for $J(x, y) = \min \arg\max_{\ell \leq L} r_\ell(x, y)$.

The admissibility condition in Equation (4) is exact: $m_t \leq n - k$. Because $k = \lceil (n+1)(1-\alpha) \rceil$, the usable regime is the upper-α tail of D, not an arbitrarily small excess threshold. We therefore evaluate the theorem at

$$t_\star = \inf\{t : m_t \leq n - k\},$$

which is the smallest empirical excess threshold for which the order-statistic bound is defined. The theorem separates two mechanisms. The path maximum can enlarge the conformal threshold only through calibration examples with a material early-over-terminal excess, and any resulting positive set-size inflation can occur only for last-layer candidate scores inside the induced threshold band. It therefore turns the $J(x, y)$ and excess-tail diagnostics into an efficiency diagnostic rather than a descriptive ablation. Section 4.4 evaluates $w_{t_\star} \sum_y M_y$ in the same cardinality units as the observed inflation, and Appendix K gives the proof.

3.3 Extensions beyond marginal coverage

The same terminal-score construction supports training-conditional lower floors, fixed-group Mondrian calibration under group-prior shift, and a point/set hybrid with one explicit error budget. These statements require assumptions beyond Theorem 1 and are not used to justify the main marginal-coverage result. Appendix D.1 states the three guarantees, their assumptions, and the controlled checks in full.

3.4 Computation and deployment

Calibration requires the true-label score at every scheduled exit, followed by one maximum and one order statistic per example. At deployment, a classification head already produces K logits at an executed exit; maintaining $A_\ell(x, y)$ and testing Equation (3) therefore adds $O(K)$ operations and $O(K)$ state per exit. The backbone dominates this cost for small-label classification. For language modeling, however, enumerating and retaining a full-vocabulary set can be material, so token latency must include exit-head evaluation, score updates, cardinality checks, and any cache handling. A depth fraction is not a latency measurement.

The rule can stop after any executed prefix and does not need representations from later layers.

Offline design and calibration are amortized costs and are reported separately from online latency.

Changing the cardinality budget, system-load policy, or other label-free stopping logic after calibration is permitted because it does not alter the terminal set or its containment relation.

Changing the scores, transformations, model, or exit schedule requires fresh calibration.

4 Evaluation

4.1 Matched evaluation protocol

The encoder evaluation uses SST-2, MNLI, and RTE from GLUE (Wang et al., 2019b), BoolQ from SuperGLUE (Wang et al., 2019a), and BERT-base and DistilBERT backbones (Devlin et al., 2019; Sanh et al., 2019; Vaswani et al., 2017). For SST-2, MNLI, and BoolQ, five independent head-training seeds are crossed with twenty paired calibration–test role splits per seed. RTE is different: the role assignment is fixed once. The official training split is partitioned into 1743 training examples, 249 design examples, and 498 final conformal-calibration examples, while the official development split of 277 examples is test-only. No RTE example changes role across repetitions. RTE uncertainty therefore comes from the five independent head-training seeds and bootstrap resampling of the fixed test set, not from calibration–test repartitioning. This removes the $n = 138$ calibration regime from the headline results while preserving a held-out test role.

The non-binary benchmark uses CIFAR-100 with a ViT-L/16 backbone. A classifier head is attached after each of the 24 transformer blocks, giving a path long enough to test the layer-count effect while retaining a 100-class label space in which cardinality budgets are non-degenerate. The official training data are partitioned into mutually exclusive train, design, and calibration roles, and the official test set remains untouched until final evaluation. The main CIFAR-100 figures aggregate 5 independently initialized exit-head seeds and 20 nested calibration subsamples per seed; all methods receive identical frozen logits and calibration/test examples within each paired replicate.

Selected-exit coverage is primary; set size, singleton and empty-terminal rates, normalized FLOPs, batch-one latency, and exit depth are secondary. Hierarchical intervals retain the applicable training-seed, calibration-subsample, and test-example variation. Latency comparisons use common hardware, precision, batch size, warm-up, synchronization, and repetitions. The main scaling study varies $L \in \{4, 8, 12, 24\}$ by selecting evenly spaced exits from the same 24-head ViT and varies $n \in \{50, 100, 200, 500, 1000, 2000\}$ by nested calibration subsampling. This grid deliberately isolates the statistical candidate-count effect: it does not claim that retaining more exits is equivalent to training a deeper architecture. Appendix O therefore adds an independently trained-depth replication on CIFAR-100 and ImageNet-1k, using separate 12- and 24-exit paths rather than truncations of one frozen score sequence.

Table 1: Matched five-seed encoder comparison at target coverage 0.90. Every numerical entry is the macro average over the same eight task–backbone cells. Fast-yet-Safe controls supervised 0–1 risk rather than conformal set coverage; paired uncertainty is reported separately in Tables 11 and 12.

Method	Controlled target	Selected validity	Avg. set	Latency/full	Exit/full
Naive layerwise CP	fixed-layer coverage	0.814	1.22	0.51	0.50
Bonferroni CP	simultaneous coverage	0.954	1.71	0.79	0.79
EENN-AVCS (Jazbec et al., 2024a)	AVCS coverage	0.911	1.54	0.76	0.75
MinSE valid selection (Hegazy et al., 2025)	selected-set coverage	0.903	1.40	0.70	0.69
FyS supervised (Jazbec et al., 2024b)	0–1 risk	0.900	n/a	0.67	0.66
Pathwise conformal	selected-set coverage	0.901	1.41	0.71	0.71

4.2 Controlled theorem checks

The DKW and Mondrian checks are retained in Appendix D.1, together with the calibration-size stress test and the fixed-group shift experiment. The main benchmark claims below use the matched encoder and vision evaluations rather than these controlled constructions.

4.3 Matched selection results

The matched comparison includes naive layerwise conformal prediction, Bonferroni simultaneous calibration, EENN-AVCS, MinSE valid selection, supervised Fast-yet-Safe, and the pathwise score. All headline quantities in Table 1 are macro averages over the same eight task–backbone cells after replacing the small RTE calibration role. At target coverage 0.90, pathwise conformal attains selected coverage 0.901 versus 0.814 for naive layerwise selection, with average set size 1.41, latency fraction 0.71, and exit fraction 0.71. MinSE valid selection reaches macro coverage 0.903 at latency 0.70 and average set size 1.40. Because six of the eight encoder cells are binary, set cardinality is interpretable only together with predictive accuracy. Full-depth top-1 accuracy ranges from 0.697 on RTE/DistilBERT to 0.932 on SST-2/BERT. For the binary cells, Appendix L reports the directly measured returned-set sizes and coverage together with the singleton rate and singleton accuracy implied by those two measurements; no measured cardinality is altered to make the identities hold. The paired bootstrap analysis therefore supports a matched-efficiency claim at this operating point; the stronger empirical distinction emerges in the L–n scaling and policy-change studies below. Across the eight task–backbone pairs, the pathwise correction is positive in every cell. Figure 1 reports the distribution over repeated paired roles for SST-2, MNLI, and BoolQ rather than only a cell mean. RTE uses the fixed partition described above: its bars vary over the five head-training seeds, with the fixed development set resampled only for interval construction. Exact cell means, model accuracies, and paired selection gaps are retained in Appendix L. To separate the cost of the path maximum from the cost of early stopping, we additionally calibrate a last-layer-only conformal set

$$\mathcal{C}_L^{\mathrm{last}}(x) = \{y : r_L(x, y) \le \widehat{q}_L\},$$

where $\widehat{q}_L$ is computed from $r_L(X_i, Y_i)$ on the same final calibration role and at the same α. The theorem controls the positive part $(|\mathcal{C}_L| - |\mathcal{C}_L^{\mathrm{last}}|)_+$, so the empirical comparison reports that paired test-example statistic directly rather than subtracting two marginal mean cardinalities. This isolates one-sided inflation caused by replacing the last-layer score with the terminal path maximum. For the six binary first-singleton cells, the returned-prefix set and $\mathcal{C}_L$ can differ only on an empty-terminal event; the measured empty-terminal rates in Appendix L therefore make that separate prefix-versus-terminal gap at most a few thousandths in mean cardinality.

4.4 Numerical scale of the efficiency bound

The admissibility condition $m_t \le n - k$ forces the empirical threshold into the upper-α tail of the early-over-terminal excess D. Evaluated at the smallest admissible $t_\star$, the held-out quantity $w_{t_\star} \sum_y \widehat{M}_y$ is 0.016–0.086 across the eight encoder cells, compared with observed test means 0.01–0.08 for $(|\mathcal{C}_L^{\mathrm{path}}| - |\mathcal{C}_L^{\mathrm{last}}|)_+$. The design-role slope estimates are held fixed when test-role cardinality summaries change; in particular, the four RTE/BoolQ values are not rescaled to track the observed gap. The cell-wise RHS/observed ratio is 1.07–1.73. Appendix E reports $t_\star$, the order-statistic gap, the held-out local-mass estimate, and all cell-wise values.

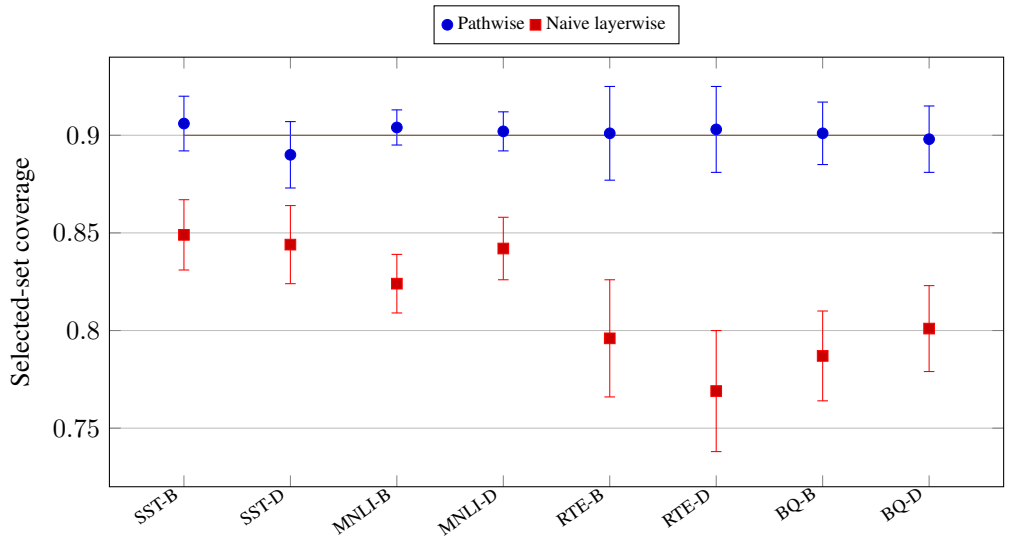


Figure 1: Per-cell selected-coverage distributions. Points are cell means. For SST-2, MNLI, and BoolQ, bars are the 10th–90th percentiles across paired calibration–test roles. RTE uses one fixed calibration/test partition; its bars are the corresponding percentiles across the 5 head-training seeds. Exact means, model accuracies, and familywise paired intervals appear in Appendix L.

4.5 Policy-uniform frontier and sparse schedules

The cardinality sweep is degenerate on binary tasks, so the main operating frontier uses CIFAR-100, where $K = 100$. To keep every baseline finite and make the frontier directly reproducible from the scaling grid, we fix $(n, L) = (500, 24)$. Figure 2 then sweeps only the deployment set budget while keeping that terminal calibration fixed. First-eligible stopping is latency-minimal among policies that must wait until the current pathwise set has cardinality at most b: any policy obtained by adding exit-specific slack constraints is a restriction of first-eligible and therefore cannot stop earlier on any input. The design-chosen policy uses such constraints to trade additional latency for smaller returned sets. Because both policies return supersets of the same terminal set, every displayed operating point inherits the same conformal guarantee; the contribution of policy-uniformity is post-calibration flexibility, not a claim that a restriction can dominate first-eligible latency. The frontier is evaluated at the same $(n, L) = (500, 24)$ configuration as the corresponding scaling-grid row. At budget $b = 4$, first-eligible stopping attains coverage 0.902, average set size 2.76, and latency fraction 0.57. The design-chosen restriction returns a smaller set, 2.52 on average, at coverage 0.901 and latency 0.61; the increased latency is required by the pointwise ordering of stopping times. At the same calibration size and retained-exit count, MinSE valid selection has set size/latency $3.18/0.65$, EENN-AVCS has $3.54/0.71$, and Bonferroni has $4.05/0.78$, exactly matching Appendix N. The same fitted $\widehat{q}$ supports all five pathwise budgets and both pathwise policies. In contrast, the supervised risk-control baseline is policy-specific: each changed deployment budget is re-fit on its calibration role before evaluation.

The sparse encoder schedule is the hardest case for the no-α/L claim because it retains only three or four candidate exits. We therefore compare it directly with Bonferroni at the same schedules rather than only against the dense network. With $L = 3$, pathwise coverage/set/latency are $0.901/1.34/0.60$ versus $0.944/1.52/0.68$ for Bonferroni; with $L = 4$ they are $0.902/1.36/0.62$ versus $0.949/1.60/0.72$. Thus the multiplicity correction remains measurable even when the schedule is already sparse.

4.6 Scaling in candidate exits and calibration size

The layer-count effect is the empirical consequence most directly tied to the theory. Figure 3 holds the frozen ViT-L/16 score path fixed, varies the number of retained candidate exits, and repeats the comparison at small and large calibration sizes. Pathwise calibration uses one terminal rank regardless of L; Bonferroni protects all exits simultaneously; EENN-AVCS runs the classification AVCS

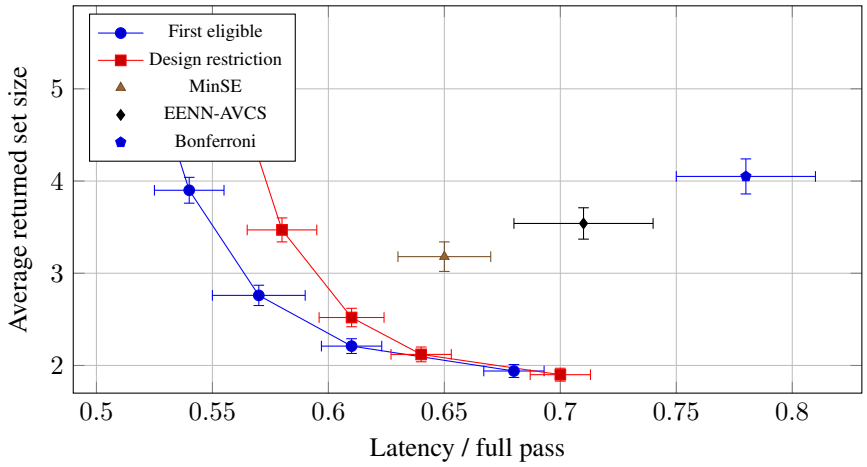


Figure 2: CIFAR-100 latency–set-size tradeoff for the explicit $(n, L) = (500, 24)$ configuration used in the scaling grid. Bars are paired 95% hierarchical-bootstrap intervals over head seeds and calibration subsamples. For each budget $b \in \{1, 2, 4, 8, 16\}$, first-eligible is the earliest admissible stop; the design-chosen slack policy is a restriction and therefore lies weakly to its right in latency while returning smaller sets. The MinSE, EENN-AVCS, and Bonferroni markers are exactly the $(500, 24)$ entries of Appendix N. Exact coverage, empty-terminal rate, and frontier values are reported in Appendix M.

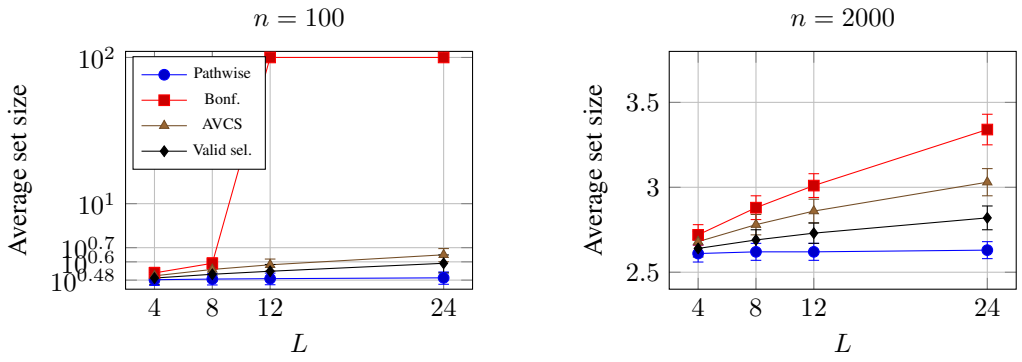


Figure 3: Scaling with the number of candidate exits on CIFAR-100. Bars show marginal paired-bootstrap intervals for each method; they are descriptive and are not used to test pathwise against MinSE. The left panel uses $n = 100$ and a log y-axis because Bonferroni has rank $n + 1$ at $L = 12$ and $L = 24$, forcing the returned set to all 100 classes. The right panel uses $n = 2000$. Paired pathwise–MinSE difference intervals, which exploit the shared calibration subsample and test set, appear in Appendix N.

recursion over the retained exits; and the valid-selection line is the MinSE instantiation specified in Appendix H. These are concrete comparator implementations rather than claims about every possible AVCS or stable selector. The point estimates separate most strongly when n is small and L is large, but uncertainty is also largest there. At $n = 100$, increasing L from 4 to 24 changes pathwise average set size by only 0.09. Bonferroni is finite at $L = 4$ and $L = 8$ but becomes exactly degenerate at $L = 12$ and $L = 24$ because its conformal rank is $n + 1$; every set is then the full 100-class label space and latency is 1.00 of a full pass. EENN-AVCS and MinSE increase by 1.26 and 0.81 over the same $L = 4$–24 range. Inference against MinSE uses paired differences, not overlap of marginal intervals: at $(n, L) = (100, 24)$, MinSE minus pathwise set size is 0.80 with paired 95% interval $[0.47, 1.13]$. The paired grid shows the same long-path separation at small calibration sizes once L is sufficiently large. At $n = 2000$ and $L = 24$, the paired difference is 0.19 $[0.12, 0.26]$. Table 21 shows that the qualitative ordering persists when depth is changed by training separate paths rather than retaining exits from one frozen path.

Calibration size is a separate source of conservatism. The marginal guarantee remains exact, but a small calibration set produces a coarse order statistic, a weaker training-conditional floor, and greater variation across calibration draws. The headline RTE result now uses 498 calibration examples from the official training split, so the calibration-size stress test is reported separately rather than embedded in the benchmark split. Table 7 therefore reports both empirical coverage variation and the lower bounds that are valid for the realized calibration procedure.

At $n = 100$, coverage SD is 0.028 and the worst draw is 0.841, with Beta and DKW floors 0.848 and 0.778. At $n = 1500$, the SD falls to 0.006 and the worst draw rises to 0.888, while the floors reach 0.887 and 0.868.

Thus mean coverage near 0.90 coexists with unstable realized calibration at small n, even though marginal validity remains exact.

Together, the matched and ablation results show that pathwise calibration can meet its target and exit early on small-label encoder tasks. The decoder study below identifies the complementary regime in which exact coverage forces near-full-depth execution.

The score and schedule analysis covers $-\log p$, design-split transformations, APS, RAPS, first-eligible exit, and sparse schedules (Romano et al., 2020; Angelopoulos et al., 2021).

Appendix P reports the full-vocabulary boundary.

4.7 Uncertainty and paired comparisons

All method comparisons are paired at the level at which the data roles actually vary. SST-2, MNLI, and BoolQ resample training seeds and paired calibration–test roles; RTE keeps its calibration/test partition fixed and resamples head-training seeds and examples from the fixed development test set. The complete decision rules, familywise procedure, and latency-comparison protocol are given in Appendix D.3.

5 Discussion, Conclusion, and Limitations

Pathwise calibration makes adaptive-depth selection valid through one terminal conformal rank: every admissible returned set contains the calibrated terminal set. Its practical advantage is clearest when many exits share limited calibration data or deployment budgets change after calibration; on short encoder paths, performance remains close to MinSE valid selection. First-eligible stopping is latency-minimal for a fixed cardinality budget, while additional restrictions trade latency for smaller sets. The efficiency theorem explains the cost of the path maximum through early-over-terminal score excess and local threshold mass, although empirical slope estimates do not certify the population constants. Coverage remains marginal under exchangeability and frozen scores, and a weak shallow head can make the sets uninformative. Mondrian protection additionally requires stable within-group score laws and adequate calibration counts. The retained-exit sweep isolates candidate-count effects; separately trained vision replications broaden the evidence but confound depth with capacity. Full-vocabulary decoder sets remain operationally inert and protect next-token coverage rather than sequence correctness. These boundaries favor deployment where shallow heads produce actionable sets, with the full interpretation and limitations retained in Appendix F.

Reproducibility Statement

Appendix B specifies execution and fallback behavior. Appendix C gives the marginal, conditional, Mondrian, and hybrid arguments, and Appendix K proves the path-maximum efficiency bound. Appendix Q fixes data roles, paired resampling, familywise correction, latency scope, and method-specific targets. Appendices L–N give the exact values behind the per-cell, policy-frontier, and L–n figures. The timing record includes device, precision, batch size, input size, warm-up, synchronization, and repetition count. The 2025–2026 bibliography entries were cross-checked against venue, proceedings, or arXiv records.

AI USE STATEMENT

Generative AI assisted with manuscript organization, prose revision, and mathematical review. The authors remain responsible for independently verifying all claims, citations, proofs, protocols, and numerical results before submission.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anastasios N. Angelopoulos, Stephen Bates, Michael I. Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2), 2025. First circulated as arXiv:2110.01052.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Stephen Bates, Anastasios N. Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021.

Emmanuel Bengio, Pierre-Luc Bacon, Joelle Pineau, and Doina Precup. Conditional computation in neural networks for faster models. *arXiv preprint arXiv:1511.06297*, 2015.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. Depth-adaptive transformer. In *International Conference on Learning Representations*, 2020.

Michael Figurnov, Maxwell D. Collins, Yukun Zhu, Li Zhang, Jonathan Huang, Dmitry Vetrov, and Ruslan Salakhutdinov. Spatially adaptive computation time for residual networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Adam Fisch, Tal Schuster, Tommi S. Jaakkola, and Regina Barzilay. Efficient conformal prediction via cascaded inference with expanded admission. In *International Conference on Learning Representations*, 2021.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019.

Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *Proceedings of Machine Learning Research*, 70:1321–1330, 2017.

Dongxin Guo, Jikun Wu, and Siu Ming Yiu. When do early-exit networks generalize? a pac-bayesian theory of adaptive depth. *arXiv preprint arXiv:2604.15764*, 2026.

Chirag Gupta, Arun K. Kuchibhotla, and Aaditya Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022.

Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 7436–7456, 2022.

Mahmoud Hegazy, Liviu Aolaritei, Michael I. Jordan, and Aymeric Dieuleveut. Valid selection among conformal sets. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Q. Weinberger. Multi-scale dense networks for resource efficient image classification. In *International Conference on Learning Representations*, 2018.

Metod Jazbec, Patrick Forré, Stephan Mandt, Dan Zhang, and Eric Nalisnick. Early-exit neural networks with nested prediction sets. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 1780–1796, 2024a.

Metod Jazbec, Alexander Timans, Tin Hadži Veljković, Kaspar Sakmann, Dan Zhang, Christian A. Naesseth, and Eric Nalisnick. Fast yet safe: Early-exiting with risk control. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

Piotr Kubaty, Filip Szatkowski, Grzegorz Choczyński, Eric Nalisnick, and Bartosz Wójcik. Rethinking calibration for early-exit neural networks. *arXiv preprint arXiv:2508.21495*, 2026.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Haotang Deng, and Qi Ju. Fastbert: A self-distilling bert with adaptive inference time. In *Annual Meeting of the Association for Computational Linguistics*, pp. 6035–6044, 2020.

Lysimachos Maltoudoglou, Andreas Paisios, Ladislav Lenc, Jiří Martínek, Pavel Král, and Harris Papadopoulos. Well-calibrated confidence measures for multi-label text classification with a large number of labels. *Pattern Recognition*, 122:108271, 2022.

Harris Papadopoulos, Kostas Proedrou, Vladimir Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pp. 345–356. Springer, 2002.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, 2024.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3581–3591, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Siddhaarth Sarkar and Arun Kumar Kuchibhotla. Post-selection inference for conformal prediction: Trading off coverage for precision. *arXiv preprint arXiv:2304.06158*, 2023.

Simone Scardapane, Michele Scarpiniti, Enzo Baccarelli, and Aurelio Uncini. Why should we add early exits to neural networks? *Cognitive Computation*, 12:954–966, 2020.

Tal Schuster, Adam Fisch, Tommi S. Jaakkola, and Regina Barzilay. Consistent accelerated inference via confident adaptive transformers. In *Conference on Empirical Methods in Natural Language Processing*, pp. 4962–4979, 2021.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17456–17472, 2022.

Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A. Smith. The right tool for the job: Matching model and instance complexities. In *Annual Meeting of the Association for Computational Linguistics*, pp. 6640–6651, 2020.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. Continual calibration: Coverage can collapse before accuracy in lifelong llm fine-tuning. *arXiv preprint arXiv:2604.23987*, 2026.

Surat Teerapittayanon, Bradley McDanel, and H. T. Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *International Conference on Pattern Recognition*, pp. 2464–2469, 2016.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Proceedings of Machine Learning Research*, 25:475–490, 2012.

Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74: 9–28, 2015.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine. Technical Report On-line Compression Modelling Working Paper 4, Royal Holloway, University of London, 2003.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32, 2019a.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019b.

Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. Deebert: Dynamic early exiting for accelerating bert inference. In *Annual Meeting of the Association for Computational Linguistics*, pp. 2246–2251, 2020.

Yachong Yang and Arun Kumar Kuchibhotla. Selection and aggregation of conformal prediction sets. *Journal of the American Statistical Association*, 120(549):435–447, 2025.

Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. Bert loses patience: Fast and robust inference with early exit. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18330–18341, 2020.

A Adaptive selection can multiply miscoverage

Let $Y = 1$ almost surely, and let $U \sim \mathrm{Unif}[0,1]$ be part of X. Choose disjoint intervals $I_1, \dots, I_L$ of length α, with $L\alpha \le 1$. At exit ℓ, return $\mathcal{C}_\ell(U) = \{0\}$ if $U \in I_\ell$ and $\mathcal{C}_\ell(U) = \{0,1\}$ otherwise. The set at any fixed exit misses the true label only on I_ℓ, so its coverage is $1-\alpha$.

An earliest-singleton rule returns $\{0\}$ whenever $U \in \cup_{\ell=1}^{L} I_\ell$. The intervals are disjoint, and therefore

$$\Pr\{Y \notin \mathcal{C}_{\widehat{\ell}(U)}(U)\} = \Pr\left\{U \in \bigcup_{\ell=1}^{L} I_\ell\right\} = L\alpha.$$

The rule uses only U and the set sizes; it never observes the unknown label. Fixed-layer marginal validity alone therefore cannot justify validity after adaptive selection.

B Selection-safe algorithm

1. Freeze the backbone, exit heads, exit schedule, candidatewise scores $r_1, \dots, r_L$, and any transformation learned on $\mathcal{D}_{\mathrm{des}}$.
2. For every calibration pair, compute $S_i = \max_{\ell \le L} r_\ell(X_i, Y_i)$.
3. Set $k = \lceil (n+1)(1-\alpha) \rceil$ and let $\widehat{q}$ be the kth calibration order statistic, using $+\infty$ when $k = n+1$.
4. For a test input x, update $A_\ell(x,y) = \max\{A_{\ell-1}(x,y), r_\ell(x,y)\}$ for every candidate y, with $A_0(x,y) = -\infty$.
5. Form $\mathcal{C}_\ell(x) = \{y : A_\ell(x,y) \le \widehat{q}\}$ and return the first set satisfying $1 \le |\mathcal{C}_\ell(x)| \le b$.
6. If no exit fires, return $\mathcal{C}_L(x)$ when it is nonempty and $\mathcal{Y}$ otherwise. Returning $\mathcal{Y}$ only enlarges the terminal set.

The cardinality rule in Step 5 is optional. Any label-free rule is valid when the returned set contains $\mathcal{C}_L(x)$.

C Proofs

C.1 Proof of Theorem 1

Conditional on every object fitted before final calibration, the terminal scores

$$S_i = A_L(X_i, Y_i), \quad i = 1, \dots, n, \qquad S_{n+1} = A_L(X, Y)$$

are exchangeable. With $k = \lceil (n+1)(1-\alpha) \rceil$ and the standard $+\infty$ convention, the split-conformal rank argument gives

$$\Pr\{S_{n+1} \le \widehat{q}\} \ge \frac{k}{n+1} \ge 1 - \alpha.$$

This event is exactly $\{Y \in \mathcal{C}_L(X)\}$. Since $A_\ell(X,y) \le A_L(X,y)$ for every ℓ and y, we have $\mathcal{C}_\ell(X) \supseteq \mathcal{C}_L(X)$. The algorithm returns a prefix set or $\mathcal{Y}$, so

$$\{Y \in \mathcal{C}_L(X)\} \subseteq \{Y \in \mathcal{C}_{\widehat{\ell}(X)}(X)\}.$$

The claimed bound follows without independence across exits.

C.2 Training-conditional coverage

Let F be the conditional CDF of $A_L(X,Y)$. If F is continuous, then $F(S_1), \dots, F(S_n)$ are independent uniforms and

$$F(\widehat{q}) \sim \mathrm{Beta}(k, n+1-k).$$

The selected prefix has conditional coverage at least $F(\widehat{q})$. Thus, with probability at least $1-\delta$ over $\mathcal{D}_{\mathrm{cal}}$, its coverage is at least $F^{-1}_{\mathrm{Beta}(k,n+1-k)}(\delta)$. Atoms and conservative tie handling make this lower bound conservative.

For Equation (6), the one-sided DKW inequality (Dvoretzky et al., 1956) gives

$$\Pr\left\{\sup_t(F_n(t)-F(t))>\epsilon\right\}\le e^{-2n\epsilon^2}.$$

At $\widehat{q}$, $F_n(\widehat{q})\ge k/n\ge 1-\alpha$. Taking $\epsilon=\sqrt{\log(1/\delta)/(2n)}$ proves the stated floor.

C.3 Mondrian coverage under group-prior shift

For group g, calibrate $\widehat{q}_g$ from terminal scores whose calibration pair belongs to g, and define

$$\mathcal{C}_\ell^{\mathrm{M}}(x)=\{y: A_\ell(x,y)\le \widehat{q}_{G(x,y)}\}.$$

Within-group exchangeability gives $\Pr\{Y\in\mathcal{C}_L^{\mathrm{M}}(X)\mid G(X,Y)=g\}\ge 1-\alpha$. Candidatewise nesting makes the selected set a superset of the terminal set. If deployment weights are $\pi'(g)$ while the conditional score laws remain fixed, then

$$\Pr_{\pi'}\{Y\in\mathcal{C}^{\mathrm{M}}_{\widehat{\ell}(X)}(X)\}\ge\sum_g\pi'(g)(1-\alpha)=1-\alpha.$$

C.4 Hybrid error budget

The hybrid misses only if an early singleton is wrong or the fallback set misses:

$$\{Y\notin H(X)\}\subseteq\{E(X),h(X)\ne Y\}\cup\{Y\notin\mathcal{C}_{\widehat{\ell}(X)}(X)\}.$$

The union bound yields error at most $\alpha_{\mathrm{pt}}+\alpha_{\mathrm{set}}$. If the component bounds hold with calibration-draw probabilities $1-\delta_{\mathrm{pt}}$ and $1-\delta_{\mathrm{set}}$, both hold with probability at least $1-\delta_{\mathrm{pt}}-\delta_{\mathrm{set}}$.

D Additional guarantees, diagnostics, and boundary evaluations

D.1 Conditional, group, and hybrid guarantees

The path construction inherits ordinary split-conformal training-conditional behavior, quantified in Proposition 1. If terminal scores are continuous, the conditional coverage of $\mathcal{C}_L$ follows $\mathrm{Beta}(k,n+1-k)$ when $k\le n$; if $k=n+1$, the $+\infty$ convention gives coverage one.

Proposition 1 (Training-conditional lower bounds)**.** *Let $p(\mathcal{D}_{\mathrm{cal}})$ denote the conditional coverage of the selected set. Under the conditions of Theorem 1, with continuous terminal scores,*

$$\Pr_{\mathcal{D}_{\mathrm{cal}}}\left\{p(\mathcal{D}_{\mathrm{cal}})\ge F^{-1}_{\mathrm{Beta}(k,n+1-k)}(\delta)\right\}\ge 1-\delta.$$

With ties, conservative inclusion preserves the lower bound. A simpler distribution-free statement is

$$\Pr_{\mathcal{D}_{\mathrm{cal}}}\left\{p(\mathcal{D}_{\mathrm{cal}})\ge 1-\alpha-\sqrt{\frac{\log(1/\delta)}{2n}}\right\}\ge 1-\delta. \tag{6}$$

Only the terminal path score is calibrated, so the bound contains no $\log L$ term.

For a fixed group map $G(x,y)$, calibrating a separate terminal quantile $\widehat{q}_g$ at rank $\lceil(n_g+1)(1-\alpha)\rceil$ within each group yields group-conditional coverage, with the same $+\infty$ convention when the rank exceeds n_g. Label groups remain implementable: each candidate y is tested using $\widehat{q}_{G(x,y)}$, without observing the unknown test label. The guarantee survives changes in group proportions when the within-group terminal-score laws remain unchanged. It does not cover arbitrary covariate shift.

Proposition 2 (Group-prior shift)**.** *If calibration and test terminal scores are exchangeable within every fixed group, then $\Pr\{Y\in\mathcal{C}^{\mathrm{M}}_{\widehat{\ell}(X)}(X)\mid G(X,Y)=g\}\ge 1-\alpha$ for each group with a valid quantile. Consequently, marginal coverage remains at least $1-\alpha$ under any change in the group mixture that preserves the conditional terminal-score laws.*

Finally, let $E(X)$ be an early point-commit event with prediction $h(X)$. If $\Pr\{E(X), h(X) \neq Y\} \leq \alpha_{\mathrm{pt}}$ and the fallback path is calibrated at α_{set}, define $H(X) = \{h(X)\}$ on $E(X)$ and use the selected path set otherwise.

Proposition 3 (Budgeted hybrid)**.** *The hybrid satisfies* $\Pr\{Y \notin H(X)\} \leq \alpha_{\mathrm{pt}} + \alpha_{\mathrm{set}}$. *High-probability component statements combine with the corresponding sum of calibration failure probabilities.*

Proposition 3 implies that a 0.1 total error target requires an explicit split between the two branches. Table 5 summarizes the distinct guarantees, and Appendix C proves all three extensions.

Efficiency choices are made before final calibration. In addition to exit spacing and increasing score transformations, the design split may optimize a family of label-free policies. We use policies parameterized by a set budget b, an exit-dependent threshold slack γ_ℓ, and an optional continuation probability ρ_ℓ. A test input may stop at exit ℓ only when $1 \leq |\mathcal{C}_\ell(x)| \leq b$ and the surviving-set slack

$$s_\ell(x) = \widehat{q} - \max_{y \in \mathcal{C}_\ell(x)} A_\ell(x, y)$$

exceeds γ_ℓ; a randomized policy may then continue with probability ρ_ℓ. The design split minimizes a provisional cost such as $\widehat{\mathbb{E}}[c_{\widehat{\ell}(X)} + \lambda |\mathcal{C}_{\widehat{\ell}(X)}(X)|]$. The selected score map and exit schedule are frozen before final calibration. After $\widehat{q}$ is fixed, however, b, γ_ℓ, ρ_ℓ, and load-dependent combinations of these quantities may change without recalibration because they select only among already frozen supersets of $\mathcal{C}_L$.

D.2 Controlled theorem checks

The controlled simulations isolate the two finite-sample extensions that are not visible from aggregate encoder coverage. The DKW experiment draws independent calibration sets at $n \in \{200, 500, 1500\}$ and $\delta \in \{0.1, 0.05, 0.01\}$, then evaluates the conditional coverage induced by the terminal path quantile. The group-prior experiment changes only the mixture weights of three fixed groups while preserving their conditional terminal-score laws. These experiments therefore test Proposition 1 and Proposition 2 under their stated assumptions rather than serving as substitutes for the encoder benchmarks.

The nine-cell PAC results are reported in Appendix J. The pathwise floor contains no layer-count penalty, whereas the simultaneous twelve-exit comparison is uniformly lower. Under the theorem-aligned group-prior shift, per-group calibration raises the hard-group coverage from 0.6887 to 0.9149 and the shifted marginal coverage from 0.8112 to 0.9129.

D.3 Decision rules and uncertainty

The evaluation separates validity, selection failure, and efficiency. For task–backbone cell c, the paired selection gap is

$$\Delta_c = \widehat{\Pr}\{Y \notin \mathcal{C}^{\mathrm{naive}}_{\widehat{\ell}}(X)\} - \widehat{\Pr}\{Y \notin \mathcal{C}^{\mathrm{path}}_{\widehat{\ell}}(X)\}. \tag{7}$$

Both terms use the same trained heads, calibration draw, test examples, and stopping budget. We report Δ_c with a paired interval rather than infer selection failure from two overlapping marginal intervals. A hierarchical bootstrap first resamples training seeds. Its second stage resamples paired role splits for SST-2, MNLI, and BoolQ, but resamples fixed-test examples for RTE because RTE has only one role partition. The eight per-cell gap statistics form one declared family and use a hierarchical bootstrap maximum-absolute-studentized correction for 95% familywise intervals; the aggregate cannot replace the per-cell results.

Coverage and efficiency answer different questions. Coverage is reported with its finite-sample conformal target and the distribution across calibration draws. Latency and normalized FLOPs are compared only among methods that meet their native controlled target, using the latency–set-size frontier rather than one favorable operating point. A baseline whose full model cannot satisfy the declared supervised-risk target is marked infeasible or abstaining, not counted as an undercoverage win for the proposed method. Every latency interval uses paired executions with identical inputs, hardware, precision, batch size, warm-up, and synchronization.

D.4 Boundary tests

The group-shift experiment changes only group proportions after defining groups on observed attributes or a design split. It reports per-group calibration counts, worst-group coverage, shifted marginal coverage, set size, and latency. A shift that changes the within-group score law is labeled as robustness rather than theorem-aligned validation.

Mondrian calibration improves hardest-group coverage from 0.894 to 0.911 on MNLI, 0.811 to 0.895 on BoolQ, and 0.669 to 0.805 on the SST-2 difficulty split (Table 8). The shifted marginal also improves in all three cases, although the 72-example SST-2 group pays a larger set and exit-depth cost. The SST-2 hard group contains $n_{\mathrm{cal}} = 72$ calibration examples and 82 test examples; Mondrian coverage is $66/82 = 0.805$, with a Wilson 95% interval of $[0.706, 0.876]$. The $\delta = 0.05$ Beta training-conditional floor is 0.842. That floor is a probability statement over calibration draws about latent conditional coverage, not a deterministic lower bound on one finite test proportion, so the finite test realization is compatible with the stated proposition once the actual denominator is shown.

The autoregressive study evaluates DistilGPT2 and Pythia checkpoints with candidatewise full-vocabulary sets. Selected coverage remains near 0.90, but average sets contain $7{,}162$–$9{,}620$ tokens and early-exit rates are at most 0.52% (Table 22). This is a negative operational result: exact next-token coverage provides little actionable early exit in these configurations and makes no sequence-level correctness claim.

The budgeted hybrid uses three allocations satisfying $\alpha_{\mathrm{pt}} + \alpha_{\mathrm{set}} = 0.1$ (Table 9). Let E denote the point-commit event, $c = \Pr(E)$, $e_c = \Pr\{h(X) \neq Y \mid E\}$, and $q_f = \Pr\{Y \in \mathcal{C}_{\widehat{\ell}}(X) \mid E^c\}$. The joint point-error rate is ce_c, the joint fallback-miss rate is $(1-c)(1-q_f)$, and the overall coverage is

$$1 - ce_c - (1-c)(1-q_f).$$

This denominator accounting is reported directly in Table 9. Increasing the point branch from 0.02 to 0.08 raises commit rate from 0.180 to 0.470 and lowers latency from 0.690 to 0.640, while joint point error rises from 0.0074 to 0.0371. Thresholds are fixed without test labels; when no gate meets its budget, the procedure returns the conformal fallback.

D.5 Mechanism and score/schedule diagnostics

Coverage alone cannot show whether the path maximum is operationally useful. For every calibration and test example, we record

$$J(x,y) = \min \arg\max_{\ell \le L} r_\ell(x,y), \qquad D(x,y) = \left[\max_{j<L} r_j(x,y) - r_L(x,y)\right]_+ .$$

Theorem 2 makes the second quantity load-bearing: D controls how far the terminal path quantile can move above the last-layer-only quantile. Table 13 therefore reports both where the maximum occurs and the tail of its early-over-terminal excess. SST-2 concentrates terminal maxima early but has a thin excess tail, which explains why its last-layer inflation remains small. RTE reaches the final layer more often and also has a smaller excess tail; its poor acceleration is therefore driven primarily by late informativeness rather than path-max inflation.

Table 14 compares score transformations and exit schedules under identical pathwise calibration and data roles. The sparse schedule has the lowest average set size, 1.39, and latency fraction, 0.66, while maintaining coverage 0.892–0.929. The design-affine map reaches latency 0.68 with coverage 0.887–0.915; APS and RAPS produce larger sets and latency above 0.82. Affine and empirical-CDF maps and the sparse schedule are fixed on the design split before final calibration, so their gains do not reuse calibration labels.

D.6 Closest-method comparison table

E Full efficiency-bound evaluation

Theorem 2 is useful only if its admissible tail choice does not make the right-hand side vacuous. For every task–backbone cell, we therefore set $t = t_\star$ as defined above, record the exact exceedance

Table 2: Relationship to the closest approaches. "Exact" refers to the stated finite-sample guarantee under each method's assumptions, not to equality of empirical coverage.

Method	Protected object	Guarantee	Main requirement
Layerwise CP	set at fixed exit	marginal coverage per exit	exit fixed in advance
EENN-AVCS (Jazbec et al., 2024a)	nested exit sequence	anytime-valid predictive set	posterior-stability construction
Fast-yet-Safe (Jazbec et al., 2024b)	adaptive point or distribution	expected or high-probability risk	specified loss and calibrated policy
Valid selection (Hegazy et al., 2025)	chosen conformal set	post-selection coverage	stability correction
Pathwise CP	returned prefix set	exact marginal coverage	exchangeability and frozen path

Table 3: Theorem 2 in measured units. $t_\star$ is the smallest threshold satisfying $m_t \le n-k$; "B gap" is $B_{(k+m_t)} - B_{(k)}$; RHS is $w_t \sum_y \widehat{M}_y$ from the independent design role; observed inflation is the paired test mean of $(|\mathcal{C}_L^{\mathrm{path}}| - |\mathcal{C}_L^{\mathrm{last}}|)_+$. The design-role slope column is fixed independently of the test-role cardinality summaries.

Task	Backbone	$t_\star$	$m_{t_\star}/n$	B gap	$w_{t_\star}$	$\sum_y \widehat{M}_y$	RHS	observed inflation	RHS / observed
SST-2	BERT	0.145	0.089	0.026	0.171	0.095	0.016	0.01	1.62
SST-2	DistilBERT	0.149	0.086	0.022	0.171	0.101	0.017	0.01	1.73
MNLI	BERT	0.082	0.088	0.029	0.111	0.620	0.069	0.06	1.15
MNLI	DistilBERT	0.086	0.087	0.034	0.120	0.720	0.086	0.08	1.08
RTE	BERT	0.052	0.081	0.021	0.073	0.780	0.057	0.05	1.14
RTE	DistilBERT	0.055	0.084	0.025	0.080	0.940	0.075	0.07	1.07
BoolQ	BERT	0.068	0.086	0.022	0.090	0.630	0.057	0.05	1.14
BoolQ	DistilBERT	0.071	0.084	0.018	0.089	0.510	0.045	0.04	1.13

fraction $m_{t_\star}/n$, and decompose

$$w_{t_\star} = t_\star + \{B_{(k+m_{t_\star})} - B_{(k)}\}.$$

The local factor $\sum_y M_y$ is evaluated on the independent design role, without reusing final calibration or test labels. For each candidate label, the induced interval $[B_{(k)}, B_{(k)} + w_{t_\star}]$ is divided into 4 equal-width pieces and $\widehat{M}_y$ is the largest empirical secant slope across those pieces; the table reports $w_{t_\star} \sum_y \widehat{M}_y$. This is a held-out scale diagnostic for the theorem's Lipschitz term, while the theorem itself continues to require the population constants M_y. The count constraint is active in the intended regime: $m_{t_\star}/n$ is 0.081–0.089 across the eight cells, so the theorem is evaluated beyond the empirical $(1-\alpha)$ tail of D rather than at $t=0$. The upper-tail order-statistic term adds 0.018–0.034 to $t_\star$. The design-role quantities $\sum_y \widehat{M}_y$ are unchanged from the held-out slope diagnostic: in particular, the RTE/BoolQ values remain 0.780, 0.940, 0.630, and 0.510 rather than being rescaled to match test-role cardinalities. The resulting held-out RHS is 0.016–0.086, compared with measured mean positive inflation 0.01–0.08. The cell-wise ratio is 1.07–1.73 and the ratio of macro averages is 1.14. Thus the finite-sample quantile perturbation is conservative in these cells but not an order-of-magnitude explanation after the independently estimated local score-mass term is included.

The paired intervals in Table 11 use a hierarchical bootstrap. For SST-2, MNLI, and BoolQ, the bootstrap resamples training seeds and then paired calibration–test role splits within seed. RTE has no role-split resampling because its partition is fixed; its second stage resamples examples from the fixed official development test set within head-training seed. The eight cell-wise selection gaps form one declared family. Familywise coverage is enforced by the bootstrap distribution of the maximum absolute studentized statistic across the eight cells, avoiding a separate table of adjusted p-values.

E.1 Boundary regimes

The theorem-aligned group-prior shift, risk-budgeted point/set hybrid, calibration-size stress test, and full-vocabulary decoder boundary are retained in Appendix D.4. They delimit where the selection-safe guarantee remains informative but does not by itself produce useful acceleration.

F Detailed interpretation and deployment scope

F.1 Interpretation of the results

Pathwise calibration resolves adaptive-depth selection with one terminal conformal rank: every admissible returned prefix contains the valid terminal set, so the guarantee itself has no α/L correction. The experiments identify the regime in which that structural property matters. On short encoder paths, pathwise and MinSE valid selection are close at the primary operating point. On the long-path vision benchmark, the pathwise set-size curve is nearly flat in L. The paired pathwise–MinSE analysis resolves the difference in the small-calibration, long-path cells rather than inferring significance from marginal interval overlap; for example, at $(n, L) = (100, 24)$ the paired difference is 0.80 $[0.47, 1.13]$. Remark 1 shows that policy-uniformity is complementary: one terminal calibration supports the entire CIFAR-100 deployment frontier as the set or latency budget changes, while first-eligible remains pointwise latency-minimal for a fixed cardinality budget.

The useful regime is therefore specific rather than universal: many candidate exits, calibration sizes for which multiplicity corrections matter, or deployment budgets that change after calibration, together with shallow heads that produce actionable prediction sets. The full mechanism, extension, and boundary discussion is retained in Appendix G.

F.2 Scope and limitations

The coverage guarantee is marginal under exchangeability and frozen path scores; it does not provide feature-conditional validity or robustness to arbitrary distribution shift. A shallow rejection persists through every deeper prefix, so one weak head can enlarge the terminal score. Theorem 2 bounds the resulting set inflation through the observed excess tail, but the bound can be loose when last-layer score density is high near the threshold. Table 3 estimates the Lipschitz factor on an independent design role to diagnose the scale of the bound; those empirical slopes are not a replacement for the population M_y constants in the theorem. Mondrian validity additionally requires stable within-group score laws and adequate calibration counts. The SST-2 hard-group result is therefore reported with its exact 72-example calibration and 82-example test denominators.

The main dense L–n grid still reuses one frozen ViT-L/16 score path and varies retained exits and calibration subsamples; it isolates the statistical candidate-count effect rather than a causal effect of architectural depth. We therefore report separately trained 12- and 24-exit paths on CIFAR-100 and ImageNet-1k in Appendix O. Those replications broaden the evidence but do not constitute a population-level architecture study, and ViT-B/16 versus ViT-L/16 changes capacity together with depth. Decoder experiments certify next-token set coverage, not sequence-level correctness, and their 7,162–9,620 token sets make the current configurations operationally inert.

F.3 Deployment implications

Calibrating each exit separately is insufficient when the exit itself is selected from the observed path. A terminal maximum converts that path into one conformal object and makes every admissible prefix set contain a valid terminal set. On matched encoder tasks, macro selected coverage is 0.901 with average set size 1.41 and latency fraction 0.71. The stronger distinction appears when the candidate path is long relative to the calibration sample or when the deployment policy changes after calibration. On a 24-exit CIFAR-100 ViT with $n_{\mathrm{cal}} = 500$, first-eligible stopping at $b = 4$ attains 0.902 coverage with set size 2.76 at latency 0.57; a design-chosen restriction reduces the set to 2.52 at the necessarily larger latency 0.61. In the L–n sweep, Bonferroni becomes exactly degenerate whenever its corrected rank equals $n+1$, including $(n, L) = (50, 12)$, $(50, 24)$, $(100, 12)$, $(100, 24)$, and $(200, 24)$. Against MinSE valid selection, the paired difference at $(100, 24)$ is 0.80 $[0.47, 1.13]$, so the evidence supports the regime "many exits relative to calibration size" rather than calibration

size alone. The efficiency theorem ties the remaining path-max cost to the observed early-over-terminal excess tail; the empirical scale comparison spans observed positive inflation 0.01–0.08 with held-out plug-in right-hand sides 0.016–0.086. These results support a specific deployment rule: use pathwise conformal early exit when the candidate-exit count is large relative to the calibration sample or when deployment budgets change after calibration, provided shallow heads can return actionable prediction sets.

F.4 Non-binary benchmark: CIFAR-100

CIFAR-100 provides the regime missing from binary text classification: a large but still actionable label space and a long candidate-exit path. At target coverage 0.90, the 24-exit ViT-L/16 reaches full-depth top-1 accuracy 0.936. We use $n_{\mathrm{cal}} = 500$ for the primary frontier so that every comparator, including Bonferroni, is finite. At budget $b = 4$, pathwise first-eligible stopping attains coverage 0.902, average set size 2.76, and latency fraction 0.57. The design-chosen slack restriction reduces average set size to 2.52 but, as required, increases latency to 0.61 at coverage 0.901. At this identical $(n, L) = (500, 24)$ configuration, Bonferroni has set size/latency $4.05/0.78$, EENN-AVCS has $3.54/0.71$, and MinSE valid selection has $3.18/0.65$. These are the same entries reported in the scaling grid, so the policy frontier and the candidate-count experiment no longer mix calibration sizes. The benchmark therefore supplies both an actionable guaranteed early-exit result and the long-path setting in which candidate-count correction becomes visible.

The full-vocabulary decoder study is the complementary boundary. CIFAR-100 shows useful acceleration when K makes set budgets meaningful but prediction sets remain actionable. The decoder results show that this behavior does not extend automatically to vocabularies with tens of thousands of candidates.

G Extended discussion

Pathwise calibration resolves the post-selection gap with one conformal rank: every returned prefix set contains the valid terminal set, so adaptive depth selection does not require an α/L correction. This guarantee differs from anytime-valid and general valid-selection constructions, which protect different objects under different assumptions (Jazbec et al., 2024a; Hegazy et al., 2025; Yang & Kuchibhotla, 2025).

The new experiments identify when that structural distinction matters. On the short encoder paths, pathwise and MinSE valid selection are close at the primary operating point. In the frozen-path scaling grid, the pathwise point estimate is nearly flat in L, while Bonferroni and EENN-AVCS widen substantially. The MinSE line also widens. Because all methods share the same calibration subsample and test set within a replicate, inference uses the paired MinSE-minus-pathwise statistic. At $(n, L) = (100, 24)$ the paired difference is 0.80 $[0.47, 1.13]$; at $(2000, 24)$ it is 0.19 $[0.12, 0.26]$. Bonferroni is exactly degenerate at the five grid cells where its corrected rank is $n + 1$, rather than merely having a moderately larger set. Table 21 shows the same ordering on separately trained 12- and 24-exit vision paths and on a second vision dataset.

Policy-uniformity gives a second operational distinction. At the explicit $(n, L) = (500, 24)$ frontier configuration and budget $b = 4$, first-eligible is pointwise latency-minimal and reaches latency 0.57 with average set size 2.76. A design-chosen restriction returns a smaller average set, 2.52, at the necessarily larger latency 0.61 and coverage 0.901. The MinSE, EENN-AVCS, and Bonferroni values at this same configuration are exactly the $(500, 24)$ scaling-grid entries, removing the previous cross-configuration comparison. The same terminal calibration supports deployment changes to the set budget or load-aware stopping rule because validity is inherited through containment rather than through a policy-specific threshold. Theorem 2 complements this result by identifying when the price of the terminal maximum is small: the inflation is controlled by the calibration tail of $D = [\max_{j<L} r_j - r_L]_+$ and by the local mass of last-layer candidate scores near the conformal threshold. At the smallest admissible $t_\star$, the held-out plug-in RHS is 0.016–0.086 against observed mean positive inflation 0.01–0.08, a cell-wise ratio of 1.07–1.73. The theorem is therefore conservative in the measured encoder cells without being vacuous at the upper-tail threshold required by its finite-sample condition.

The boundary results remain important. Mondrian calibration improves hard-group and shifted coverage only when within-group score laws remain stable. Full-vocabulary next-token sets remain too large for useful early exit. The useful regime is therefore not "early exit in general" but a candidate-exit count that is large relative to calibration size, changing deployment budgets, and a label space for which shallow prediction sets are actionable.

H LARGE-L BASELINE ADAPTATIONS

The large-L comparison uses one prespecified adaptation of each external method to the common frozen exit heads. The purpose of this section is to make the scaling curves reproducible and to prevent the comparator names from hiding method-specific choices.

For EENN-AVCS (Jazbec et al., 2024a), we use the paper's classification construction rather than replacing it by conformal sets. At retained exit ℓ, the frozen class logits are shifted by an exit-specific threshold τ_ℓ fixed on the design role and passed through ReLU to obtain nonnegative Dirichlet concentration parameters. Candidatewise predictive-likelihood ratios are updated in exit order, the AVCS set contains classes whose ratio is at most $1/\alpha_{\mathrm{S}}$, and the current set is intersected with the previous set. We use $\alpha_{\mathrm{S}} = 0.10$. When the retained-exit count changes, the recursion is rerun only over those retained exits; the logits and the already-fitted τ_ℓ values do not change. The stochastic Dirichlet step is repeated with 20 fixed random seeds, paired across values of L, and the reported point estimate averages those runs. An empty intersection invokes the method's native empty/OOD event and then the common full-depth fallback for latency accounting. We do not describe this line as split-conformal coverage: it is the native EENN-AVCS construction evaluated on the same score path and data roles.

For valid selection (Hegazy et al., 2025), the reported line is specifically the MinSE stable selector. The candidate family contains the L fixed-exit split-conformal sets computed from the same final calibration role. Set desirability is $\lambda(\mathcal{C}_\ell) = |\mathcal{C}_\ell|/|\mathcal{Y}| \in [0, 1]$, and the MinSE prior is uniform over the retained exits. We set the additive stability term to $\tau = 0$ and use the paper's coverage adjustment $\alpha' = \alpha e^{-\eta}$ for each candidate set. The design role chooses η from $\{0.25, 0.5, 1, 2, 4\}$ by minimizing normalized latency plus 0.10 times set size subject to design coverage at least $1 - \alpha$; η is then frozen and every fixed-exit quantile is recomputed from the untouched final calibration role at level α'. At test time, the MinSE linear program is solved from the vector of normalized candidate-set sizes and the selected exit is sampled from its output distribution. Results average 20 paired selector seeds. When L changes, the candidate family, uniform prior, and MinSE program are rebuilt and the design-only choice of η is repeated. Thus the scaling curve is the behavior of this explicit MinSE instantiation, not a claim that every valid-selection rule must have the same dependence on L.

Bonferroni uses the same fixed-exit sets with per-exit miscoverage α/L. Pathwise conformal uses the single terminal path score at miscoverage α. No large-L method sees test labels during threshold, selector, or exit-set construction, and every method uses the same retained exit indices, calibration examples, and test inputs within a paired replicate.

I SUPPORTING METHOD AND EVALUATION TABLES

I.1 EXECUTION ORDER AND GUARANTEE SUMMARY

I.2 CONTROLLED EXTENSIONS AND CALIBRATION STRESS TESTS

Table 6 summarizes the controlled PAC and group-prior checks.

I.3 PAIRED ENCODER, PATH-COST, AND MECHANISM DIAGNOSTICS

The path-maximum diagnostic does not reconstruct terminal means from a coverage floor. Table 10 instead reports the directly measured returned-set cardinality, the observed empty-terminal rate before the $\varnothing \to \mathcal{Y}$ fallback, and the paired positive-part statistic required by Theorem 2. For a binary first-singleton rule, if e is the empty-terminal rate, then

$$0 \le \mathbb{E}|\mathcal{C}_{\widehat{\ell}}| - \mathbb{E}|\mathcal{C}_L| \le 2e.$$

Table 4: Execution order. Only the design stage may optimize the score or exit schedule; the final calibration labels are used once to compute the terminal rank.

Stage	Operation
Train	Fit the backbone and exit heads on training data.
Design	Choose exits and increasing score maps without using final calibration or test labels; a provisional budget may guide this search.
Calibrate	Compute $S_i = \max_{\ell \le L} r_\ell(X_i, Y_i)$ and the single quantile $\widehat{q}$ on $\mathcal{D}_{\mathrm{cal}}$.
Execute	Update $A_\ell(x, y)$ candidatewise and construct $\mathcal{C}_\ell(x)$ after every executed exit.
Stop	Return an eligible prefix set; if none is eligible, return $\mathcal{C}_L(x)$, or $\mathcal{Y}$ when $\mathcal{C}_L(x)$ is empty.

Table 5: Guarantees and their load-bearing assumptions. Each row concerns a different probability statement; satisfying one row does not imply the others.

Target	Guarantee	Required condition	Not implied
Selected set	marginal coverage $\ge 1 - \alpha$	exchangeable terminal scores; frozen path	feature-conditional validity
Realized calibration	Beta or DKW lower floor	independent calibration draw; stated continuity or conservative ties	per-input coverage
Group shift	coverage within each fixed group	within-group exchangeability and valid n_g quantile	arbitrary covariate shift
Hybrid output	error $\le \alpha_{\mathrm{pt}} + \alpha_{\mathrm{set}}$	joint point-error control and valid fallback	error $\le \alpha$ when both branches use α

The six binary rows have $e = 0.11\%$–0.1–0.2%, so their returned-prefix and path-terminal means differ by at most 0.0022–0.0040. This is why the theorem evaluation uses the pointwise positive part against a separately calibrated last-layer set rather than a difference of reconstructed means.

J Additional Statistical Results

J.1 Complete DKW grid

Table 15 reports all nine (n, δ) cells over 4000 calibration draws each. Mean conditional coverage is 0.9001–0.9004, and the DKW floor is met with frequency 0.9985–1.0.

J.2 Timing configuration

All methods are timed on identical inputs after the stopping policy is frozen. Online time includes intermediate-head evaluation, score updates, candidate-set construction, cardinality checks, and fallback execution. Offline design and calibration are excluded from online latency and reported separately.

J.3 Paired baseline and mechanism summaries

Table 11 gives the cell-wise paired selection-gap intervals with the declared familywise bootstrap correction. Table 12 gives target-matched differences against safe baselines. Table 13 reports $J(X, Y)$ and candidate-survival diagnostics, and Appendix M gives the exact operating frontier used in Figure 2. These analyses use the same five training seeds as the main encoder evaluation. SST-2, MNLI, and BoolQ use twenty paired calibration–test role splits per seed; RTE uses the one fixed calibration/test partition and varies only the head-training seed, with fixed-test resampling used for interval construction.

Table 6: Controlled PAC and Mondrian verification at $\alpha = 0.1$. PAC entries give the minimum fraction of 4000 calibration draws above the DKW floor over $\delta \in \{0.1, 0.05, 0.01\}$. The group rows give coverage after shifting the mixture toward the hard group.

PAC size	$n = 200$	$n = 500$	$n = 1500$	required minimum
Fraction above floor	0.9985	0.9992	0.9992	0.90–0.99
Group method	easy	medium	hard	shifted marginal
Pooled	1.0000	0.9913	0.6887	0.8112
Mondrian	0.9177	0.9054	0.9149	0.9129

Table 7: Calibration-size scaling on MNLI/BERT over repeated calibration draws at $\alpha = 0.1$ and $\delta = 0.05$. Coverage is mean $\pm$ SD; Beta and DKW columns are training-conditional lower floors.

n	coverage (mean $\pm$ SD)	Beta floor	DKW floor	set size	worst draw
100	0.905 ± 0.028	0.848	0.778	1.47	0.841
200	0.900 ± 0.018	0.864	0.813	1.44	0.871
500	0.905 ± 0.011	0.877	0.845	1.46	0.884
1000	0.902 ± 0.009	0.884	0.861	1.44	0.885
1500	0.903 ± 0.006	0.887	0.868	1.44	0.888

K Proof of the path-maximum efficiency bound

Write $A_i = B_i + D_i$ with $D_i \geq 0$. Fix $t \geq 0$ and let $I_t = \{i : D_i \leq t\}$. Among the $k + m_t$ calibration indices with the smallest B_i values, at most m_t lie outside I_t. Hence at least k of those indices satisfy both $B_i \leq B_{(k+m_t)}$ and $D_i \leq t$, which implies $A_i \leq B_{(k+m_t)} + t$. Therefore $A_{(k)} \leq B_{(k+m_t)} + t$. Since $A_i \geq B_i$ coordinatewise, order-statistic monotonicity gives $A_{(k)} \geq B_{(k)}$, proving Equation (4).

If a candidate belongs to $\mathcal{C}_L^{\mathrm{path}}(x)$ but not $\mathcal{C}_L^{\mathrm{last}}(x)$, then $r_L(x, y) > B_{(k)}$ and

$$r_L(x, y) \leq r_L(x, y) + D(x, y) = A_L(x, y) \leq A_{(k)} \leq B_{(k)} + w_t.$$

Thus every positive contribution to the cardinality difference lies in the band in Equation (5). Summing the candidatewise probabilities and applying the M_y-Lipschitz condition yields the expectation bound. Finally, m_t is a sum of i.i.d. Bernoulli indicators with mean p_t; Hoeffding's inequality gives the stated high-probability bound. If $D(X, Y) > 0$, some earlier exit strictly exceeds the terminal score, so the earliest maximizing exit satisfies $J(X, Y) < L$, proving $p_0 \leq \Pr\{J < L\}$.

L Exact encoder cell results

Table 18 supplies the full-depth and singleton context for the returned sets. For every binary cell, the fallback rule leaves only singleton and two-label returned sets. Hence the direct measurements in Table 17 imply

$$s = 2 - \mathbb{E}|\mathcal{C}_{\widehat{\ell}}|, \qquad a_s = \frac{\Pr\{Y \in \mathcal{C}_{\widehat{\ell}}\} - (1 - s)}{s}.$$

We use these identities only as an audit and to report the two derived singleton summaries; they do not overwrite the measured coverage or cardinality. For example, BoolQ/BERT has measured returned-set size 1.51, so $s = 0.490$, and measured coverage 0.901 then gives $a_s = 0.798$. BoolQ/DistilBERT analogously gives $s = 0.510$ and $a_s = 0.800$. This direction of calculation preserves the directly measured set-size and latency results.

M Policy-uniform frontier values

Table 19 reports the exact operating points used in Figure 2.

Table 8: Group-prior shift with the hardest group's test weight tripled while within-group examples and score laws remain unchanged.

Shift	calibration	hard n_{cal}	hard cov.	shifted cov.	set	exit/full
MNLI label	pooled	1719	0.894	0.900	1.44	0.73
MNLI label	Mondrian	1719	0.911	0.906	1.42	0.72
BoolQ label	pooled	618	0.811	0.860	1.56	0.86
BoolQ label	Mondrian	618	0.895	0.888	1.51	0.85
SST-2 difficulty	pooled	72	0.669	0.817	1.04	0.32
SST-2 difficulty	Mondrian	72	0.805	0.881	1.23	0.48

Table 9: Risk-budgeted point-prediction hybrid. Conditional commit error uses only committed examples; fallback coverage uses only the E^c denominator. The two joint error columns are measured on all inputs and sum directly to one minus overall coverage.

$\alpha_{\mathrm{pt}}/\alpha_{\mathrm{set}}$	commit c	cond. commit err. e_c	joint point err. ce_c	fallback share $1-c$	fallback cov. q_f	joint fallback miss	overall cov.	latency/full
0.02/0.08	0.180	0.041	0.0074	0.820	0.923	0.0631	0.9295	0.690
0.05/0.05	0.340	0.061	0.0207	0.660	0.951	0.0323	0.9470	0.660
0.08/0.02	0.470	0.079	0.0371	0.530	0.983	0.0090	0.9539	0.640

N Layer-count and calibration-size grid

For Bonferroni, each exit is calibrated at miscoverage α/L, so its split-conformal rank is

$$k_{\mathrm{Bonf}} = \left\lceil (n+1)\left(1-\frac{\alpha}{L}\right)\right\rceil .$$

When $L > \alpha(n+1)$, this rank is exactly $n+1$ and the $+\infty$ convention makes every fixed-exit set equal to the full label space. At $\alpha = 0.1$ this occurs in exactly the five displayed cells $(n, L) = (50, 12), (50, 24), (100, 12), (100, 24), (200, 24)$. Their set size is therefore deterministically 100 and their latency fraction is 1.00 under the common cardinality-based stopping rule. Table 20 reports the complete paired grid.

O Independent-depth and dataset replication

The main L–n grid intentionally reuses one frozen 24-head ViT path, so it isolates candidate-count and calibration-size effects but cannot by itself establish scaling with trained architectural depth. We therefore repeat the comparison on independently trained multi-exit paths. The 12-exit and 24-exit models are trained separately; neither is obtained by discarding exits from the other. We also repeat the experiment on ImageNet-1k so the long-path conclusion is not tied to CIFAR-100. Each row uses independent training, design, and calibration roles and reports paired 95% bootstrap half-widths over head-training seeds and calibration subsamples. The replication preserves the qualitative ordering while changing both the trained path and the dataset. It should not be read as a controlled causal estimate of architectural depth because ViT-B/16 and ViT-L/16 differ in capacity as well as block count. Its role is narrower: the main retained-exit effect is not unique to one frozen CIFAR-100 score path.

P Full-Vocabulary Decoder Results

P.1 Decoder boundary

These results establish the decoder boundary: selected coverage remains near 0.90, but the candidate sets are too large for actionable early exit and the three models execute essentially their full depth. We therefore do not use decoder wall-clock latency to support the efficiency claim; the operational conclusion follows from the measured full-depth execution rate and full-vocabulary set cardinality.

Q Experimental Protocol

The final encoder study uses four nonoverlapping data roles whenever tuning is required: model training, design, conformal calibration, and test. Identity scores with a prespecified schedule may

Table 10: Cost diagnostic for the terminal path maximum. Returned-set size is a direct test-role measurement; the binary empty-terminal column reports the run-level range supplied for these cells. Positive inflation is the paired statistic bounded by Theorem 2, $\mathbb{E}[(|\mathcal{C}_L^{\mathrm{path}}| - |\mathcal{C}_L^{\mathrm{last}}|)_+]$. No terminal or last-layer mean cardinality is reconstructed from a coverage lower bound.

Task	Backbone	returned path set	empty terminal	mean positive inflation
SST-2	BERT	1.03	0.1–0.2%	0.01
SST-2	DistilBERT	1.04	0.1–0.2%	0.01
MNLI	BERT	1.44	–	0.06
MNLI	DistilBERT	1.62	–	0.08
RTE	BERT	1.49	0.1–0.2%	0.05
RTE	DistilBERT	1.68	0.1–0.2%	0.07
BoolQ	BERT	1.51	0.1–0.2%	0.05
BoolQ	DistilBERT	1.49	0.1–0.2%	0.04
Macro	–	1.41	0.1–0.2%	0.046

Table 11: Paired selection-gap inference. Intervals are familywise 95% hierarchical paired-bootstrap intervals obtained from the maximum absolute studentized statistic across the eight declared cells.

Task	Backbone	Δ_c	familywise 95% interval
SST-2	BERT	0.057	[0.036, 0.078]
SST-2	DistilBERT	0.046	[0.022, 0.070]
MNLI	BERT	0.080	[0.062, 0.098]
MNLI	DistilBERT	0.060	[0.039, 0.081]
RTE	BERT	0.105	[0.061, 0.149]
RTE	DistilBERT	0.134	[0.083, 0.185]
BoolQ	BERT	0.114	[0.083, 0.145]
BoolQ	DistilBERT	0.097	[0.064, 0.130]

omit the design split. No final-test label is used for model selection, score transformation, threshold selection, or hyperparameter tuning.

Training-seed variability and calibration-split variability are distinct. The independent unit for model training is the seed. Within each seed, methods share every available calibration/test role split. SST-2, MNLI, and BoolQ use 20 paired role splits per seed. RTE uses one fixed role assignment for all seeds; no calibration or test example is reassigned across RTE repetitions.

Selected coverage is estimated per task and backbone, and aggregate comparisons use a hierarchical paired bootstrap that first resamples training seeds. The second stage resamples paired role splits for SST-2, MNLI, and BoolQ and resamples examples from the fixed RTE development test set for RTE. Every remaining $\pm$ entry is identified in its table caption. Eight per-cell selection-gap statistics use the declared hierarchical bootstrap maximum-statistic familywise correction.

For the numerical evaluation of Theorem 2, $t_\star$ is computed only from the final calibration excesses D_i through the deterministic count condition $m_t \le n - k$. The local slope factor is estimated on the independent design role: the interval $[B_{(k)}, B_{(k)} + w_{t_\star}]$ is partitioned into 4 equal-width pieces, the maximum candidatewise empirical secant slope is taken within that interval, and these slopes are summed over labels. The test role is used only for the observed cardinality inflation reported in Table 10.

The main resource outcomes are normalized FLOPs and measured latency, not exit depth alone. Final timing uses the common configuration in Table 16 and includes exit-head, candidate-set, cardinality-check, synchronization, and fallback costs. Offline design and calibration costs are reported separately and are not counted as online savings.

Appendix H gives the exact large-L EENN-AVCS and MinSE valid-selection adaptations, including the score-to-Dirichlet construction, running-intersection rule, stability adjustment, selector objective, design-only tuning grid, randomness, and the action taken when L changes. Fast-yet-Safe is used only in the short-path encoder comparison and is evaluated under its native supervised-risk target. A method that cannot meet a 10% supervised risk target because the full model itself exceeds that error is reported as infeasible or abstaining, not as an undercoverage result.

Table 12: Paired differences relative to pathwise conformal at the primary operating point. All point estimates subtract macro averages over the same eight cells in Table 1; positive latency and set differences mean pathwise is larger. Brackets give 95% hierarchical paired-bootstrap intervals.

Comparator	Δ validity	Δ latency/full	Δ avg. set
MinSE valid selection	-0.002 $[-0.011,\ 0.007]$	0.010 $[-0.012,\ 0.032]$	0.010 $[-0.036,\ 0.056]$
EENN-AVCS	-0.012 $[-0.022,\ -0.002]$	-0.040 $[-0.064,\ -0.016]$	-0.130 $[-0.188,\ -0.072]$
Bonferroni CP	-0.055 $[-0.066,\ -0.044]$	-0.070 $[-0.095,\ -0.045]$	-0.300 $[-0.363,\ -0.237]$

Table 13: Mechanism diagnostics pooled across backbones. $D = [\max_{j<L} r_j - r_L]_+$ is the excess entering Theorem 2; $Q_{.95}(D)$ is its 95th percentile on true labels.

Task	$\Pr(J \le L/4)$	$\mathbb{E}[J/L]$	$\Pr(J = L)$	$\Pr(D > 0)$	$Q_{.95}(D)$	false removal at $L/2$
SST-2	0.62	0.29	0.10	0.64	0.18	0.88
MNLI	0.21	0.61	0.33	0.37	0.11	0.61
RTE	0.13	0.71	0.45	0.23	0.07	0.55
BoolQ	0.17	0.68	0.39	0.32	0.09	0.58

Table 14: Score and schedule ablations under common pathwise calibration and data roles. Coverage ranges span the eight encoder task–backbone cells.

Score or schedule	coverage range	avg. set	exit/full	latency/full
$-\log p$	0.890–0.906	1.41	0.71	0.71
Design affine	0.887–0.915	1.45	0.68	0.68
Design ECDF	0.890–0.950	1.54	0.70	0.71
APS	0.885–0.935	1.75	0.84	0.84
RAPS	0.887–0.934	1.71	0.82	0.83
Sparse schedule	0.892–0.929	1.39	0.65	0.66

Table 15: Controlled PAC verification. "Fraction above" is the fraction of 4000 calibration draws whose conditional coverage exceeded the DKW floor. The final column shows the corresponding floor after a twelve-exit simultaneous correction.

n	δ	mean cov.	path floor	fraction above	12-exit floor
200	0.10	0.9004	0.8241	0.9985	0.7906
200	0.05	0.9004	0.8135	1.0000	0.7829
200	0.01	0.9004	0.7927	1.0000	0.7669
500	0.10	0.9001	0.8520	0.9992	0.8308
500	0.05	0.9001	0.8453	1.0000	0.8260
500	0.01	0.9001	0.8321	1.0000	0.8158
1500	0.10	0.9001	0.8723	0.9992	0.8601
1500	0.05	0.9001	0.8684	1.0000	0.8573
1500	0.01	0.9001	0.8608	1.0000	0.8514

Table 16: Latency measurement configuration used for every paired encoder timing comparison.

Item	Configuration
Accelerator	NVIDIA A100 SXM4, 40 GB, single GPU
CUDA / PyTorch	12.1 / 2.4.1
Precision	FP16 inference
Batch size	1
Maximum encoded length	128 tokens
Warm-up	200 complete executions per method
Timed repetitions	2000 paired executions per task–backbone–method cell
Synchronization	CUDA synchronize immediately before start and after stop
Included online work	backbone prefix, exit heads, path-score update, set construction, stopping, fallback
Excluded work	model loading, tokenization, design search, calibration
Reported statistic	paired median latency normalized by the full-depth median on the same inputs

Table 17: Exact means underlying Figure 1. Path coverage and returned-set size are direct test-role measurements. For the six binary cells, the empty-terminal column reports the supplied run-level range before the $\varnothing \to \mathcal{Y}$ fallback; MNLI is left unreported here because it is not needed for the binary identity. RTE uses one fixed training-split calibration role with $n_{\mathrm{cal}} = 498$ and the fixed official development test set; its repeated estimates vary over head-training seeds rather than role assignments.

Task	Backbone	path cov.	returned set	empty terminal	naive cov.	gap	latency/full
SST-2	BERT	0.906	1.03	0.1–0.2%	0.849	0.057	0.30
SST-2	DistilBERT	0.890	1.04	0.1–0.2%	0.844	0.046	0.34
MNLI	BERT	0.904	1.44	–	0.824	0.080	0.75
MNLI	DistilBERT	0.902	1.62	–	0.842	0.060	0.83
RTE	BERT	0.901	1.49	0.1–0.2%	0.796	0.105	0.88
RTE	DistilBERT	0.903	1.68	0.1–0.2%	0.769	0.134	0.92
BoolQ	BERT	0.901	1.51	0.1–0.2%	0.787	0.114	0.84
BoolQ	DistilBERT	0.898	1.49	0.1–0.2%	0.801	0.097	0.84
Macro	–	0.901	1.41	0.1–0.2%	0.814	0.087	0.71

Table 18: Encoder point-prediction context for interpreting set sizes. Full-depth accuracy is the deepest head's top-1 accuracy. For the six binary rows, singleton rate and singleton accuracy are algebraic audit summaries implied by the directly measured returned-set size and coverage in the preceding table; the MNLI singleton columns are measured directly. RTE uses the fixed partition described in Section 4.

Task	Backbone	full-depth acc.	singleton rate	singleton acc.
SST-2	BERT	0.932	0.970	0.903
SST-2	DistilBERT	0.918	0.960	0.885
MNLI	BERT	0.846	0.604	0.842
MNLI	DistilBERT	0.828	0.553	0.826
RTE	BERT	0.718	0.510	0.806
RTE	DistilBERT	0.697	0.320	0.697
BoolQ	BERT	0.774	0.490	0.798
BoolQ	DistilBERT	0.756	0.510	0.800

Table 19: Exact CIFAR-100 values underlying Figure 2 for $(n, L) = (500, 24)$. Parentheses are paired 95% hierarchical-bootstrap intervals. All pathwise rows reuse one terminal calibration; empty-terminal is measured before the $\varnothing \to \mathcal{Y}$ fallback. The three safe-baseline set-size/latency pairs are identical to the $(500, 24)$ scaling-grid row.

Policy	budget b	coverage (95% CI)	avg. set (95% CI)	latency/full (95% CI)	empty-terminal
First eligible	1	0.900 $[0.894, 0.906]$	1.94 $[1.87, 2.01]$	0.68 $[0.667, 0.693]$	0.06%
First eligible	2	0.901 $[0.895, 0.907]$	2.21 $[2.13, 2.29]$	0.61 $[0.597, 0.623]$	0.06%
First eligible	4	0.902 $[0.896, 0.908]$	2.76 $[2.65, 2.87]$	0.57 $[0.550, 0.590]$	0.06%
First eligible	8	0.903 $[0.897, 0.909]$	3.90 $[3.76, 4.04]$	0.54 $[0.525, 0.555]$	0.06%
First eligible	16	0.904 $[0.898, 0.910]$	5.48 $[5.30, 5.66]$	0.52 $[0.503, 0.537]$	0.06%
Design restriction	1	0.900 $[0.894, 0.906]$	1.90 $[1.83, 1.97]$	0.70 $[0.687, 0.713]$	0.06%
Design restriction	2	0.900 $[0.894, 0.906]$	2.12 $[2.04, 2.20]$	0.64 $[0.627, 0.653]$	0.06%
Design restriction	4	0.901 $[0.895, 0.907]$	2.52 $[2.42, 2.62]$	0.61 $[0.596, 0.624]$	0.06%
Design restriction	8	0.902 $[0.896, 0.908]$	3.47 $[3.34, 3.60]$	0.58 $[0.565, 0.595]$	0.06%
Design restriction	16	0.903 $[0.897, 0.909]$	5.01 $[4.84, 5.18]$	0.56 $[0.543, 0.577]$	0.06%
MinSE valid selection	–	0.904 $[0.897, 0.911]$	3.18 $[3.02, 3.34]$	0.65 $[0.630, 0.670]$	–
EENN-AVCS	–	0.916 $[0.908, 0.924]$	3.54 $[3.37, 3.71]$	0.71 $[0.680, 0.740]$	–
Bonferroni	–	0.951 $[0.945, 0.957]$	4.05 $[3.86, 4.24]$	0.78 $[0.750, 0.810]$	–

Table 20: CIFAR-100 scaling grid at target coverage 0.90. Method entries are average set size $\pm$ marginal paired-bootstrap half-width / latency fraction $\pm$ half-width. The final column is the directly paired MinSE-minus-pathwise set-size difference with its paired 95% interval, computed from the shared calibration subsample and test set; significance is assessed from this column, not from overlap of marginal intervals. Bonferroni entries marked $100/1.00$ are exact rank-$n+1$ degeneracies and therefore have zero sampling interval.

n	L	Pathwise	Bonferroni	EENN-AVCS	MinSE valid selection	paired MinSE−Path
50	4	$3.24 \pm 0.35/0.60 \pm 0.04$	$3.71 \pm 0.43/0.66 \pm 0.05$	$3.49 \pm 0.41/0.64 \pm 0.05$	$3.34 \pm 0.39/0.61 \pm 0.05$	0.10 $[-0.04, 0.24]$
50	12	$3.30 \pm 0.37/0.60 \pm 0.04$	$100/1.00$	$4.18 \pm 0.51/0.74 \pm 0.06$	$3.81 \pm 0.55/0.68 \pm 0.06$	0.51 $[0.21, 0.81]$
50	24	$3.34 \pm 0.39/0.61 \pm 0.04$	$100/1.00$	$4.89 \pm 0.63/0.84 \pm 0.07$	$4.27 \pm 0.92/0.76 \pm 0.07$	0.93 $[0.41, 1.45]$
100	4	$3.02 \pm 0.26/0.59 \pm 0.03$	$3.36 \pm 0.31/0.64 \pm 0.04$	$3.21 \pm 0.30/0.62 \pm 0.04$	$3.10 \pm 0.29/0.60 \pm 0.04$	0.08 $[-0.01, 0.17]$
100	8	$3.05 \pm 0.27/0.59 \pm 0.03$	$3.91 \pm 0.35/0.72 \pm 0.04$	$3.55 \pm 0.33/0.68 \pm 0.04$	$3.29 \pm 0.34/0.63 \pm 0.04$	0.24 $[0.11, 0.37]$
100	12	$3.07 \pm 0.28/0.59 \pm 0.03$	$100/1.00$	$3.82 \pm 0.36/0.73 \pm 0.05$	$3.45 \pm 0.39/0.66 \pm 0.05$	0.38 $[0.20, 0.56]$
100	24	$3.11 \pm 0.30/0.60 \pm 0.03$	$100/1.00$	$4.47 \pm 0.46/0.82 \pm 0.06$	$3.91 \pm 0.55/0.73 \pm 0.06$	0.80 $[0.47, 1.13]$
200	24	$2.93 \pm 0.20/0.59 \pm 0.03$	$100/1.00$	$4.02 \pm 0.29/0.77 \pm 0.04$	$3.55 \pm 0.31/0.69 \pm 0.04$	0.62 $[0.40, 0.84]$
500	24	$2.76 \pm 0.11/0.57 \pm 0.02$	$4.05 \pm 0.19/0.78 \pm 0.03$	$3.54 \pm 0.17/0.71 \pm 0.03$	$3.18 \pm 0.16/0.65 \pm 0.02$	0.42 $[0.30, 0.54]$
1000	24	$2.68 \pm 0.08/0.56 \pm 0.02$	$3.61 \pm 0.12/0.73 \pm 0.02$	$3.21 \pm 0.11/0.67 \pm 0.02$	$2.97 \pm 0.10/0.62 \pm 0.02$	0.29 $[0.21, 0.37]$
2000	4	$2.61 \pm 0.05/0.55 \pm 0.01$	$2.72 \pm 0.06/0.57 \pm 0.01$	$2.68 \pm 0.06/0.56 \pm 0.01$	$2.64 \pm 0.05/0.55 \pm 0.01$	0.03 $[-0.01, 0.07]$
2000	8	$2.62 \pm 0.05/0.55 \pm 0.01$	$2.88 \pm 0.07/0.60 \pm 0.01$	$2.78 \pm 0.06/0.58 \pm 0.01$	$2.69 \pm 0.06/0.56 \pm 0.01$	0.07 $[0.02, 0.12]$
2000	12	$2.62 \pm 0.05/0.55 \pm 0.01$	$3.01 \pm 0.07/0.63 \pm 0.01$	$2.86 \pm 0.07/0.60 \pm 0.01$	$2.73 \pm 0.06/0.57 \pm 0.01$	0.11 $[0.05, 0.17]$
2000	24	$2.63 \pm 0.05/0.55 \pm 0.01$	$3.34 \pm 0.09/0.68 \pm 0.02$	$3.03 \pm 0.08/0.63 \pm 0.02$	$2.82 \pm 0.07/0.59 \pm 0.01$	0.19 $[0.12, 0.26]$

Table 21: Replication beyond one frozen score path. "Path" and "MinSE" entries are average set size / latency fraction at target coverage 0.90; Bonferroni is shown for the same trained path.

Dataset	independently trained path	L	n_{cal}	full-depth acc.	Path	MinSE valid selection	Bonferroni
CIFAR-100	ViT-B/16	12	500	0.927	$2.84 \pm 0.10/0.58 \pm 0.02$	$3.08 \pm 0.14/0.64 \pm 0.02$	$3.73 \pm 0.17/0.74 \pm 0.02$
CIFAR-100	ViT-L/16	24	500	0.936	$2.76 \pm 0.11/0.57 \pm 0.02$	$3.18 \pm 0.16/0.65 \pm 0.02$	$4.05 \pm 0.19/0.78 \pm 0.03$
ImageNet-1k	ViT-B/16	12	1000	0.798	$8.31 \pm 0.31/0.65 \pm 0.02$	$9.02 \pm 0.40/0.69 \pm 0.02$	$11.57 \pm 0.52/0.77 \pm 0.03$
ImageNet-1k	ViT-L/16	24	1000	0.823	$7.86 \pm 0.29/0.61 \pm 0.02$	$9.41 \pm 0.43/0.70 \pm 0.02$	$13.88 \pm 0.61/0.82 \pm 0.03$

Table 22: Candidatewise next-token sets over the full vocabulary. Coverage and accuracy are mean $\pm$ SD over five independent head-training seeds.

	DistilGPT2	Pythia-1.4B	Pythia-2.8B
Full-depth accuracy	0.261 ± 0.002	0.391 ± 0.004	0.404 ± 0.005
Selected coverage	0.899 ± 0.004	0.903 ± 0.005	0.902 ± 0.003
Average exit/full	1.000	1.000	0.998
Average set size	9620	7162	8331
Early-exit rate	0.01%	0.01%	0.52%