A REGIME THEORY OF CONTROLLER CLASS SELECTION FOR LLM ACTION DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language and vision-language models must decide whether to answer directly, retrieve evidence, defer to a stronger model, or abstain. Existing routing methods refine routers of a preselected type but offer little guidance on choosing the type itself. We address this gap through controller class selection, where a controller maps inputs to actions. We organize controllers into four classes: fixed actions, partition routers, instance-level learned controllers, and prior-gated controllers. With finite data, greater flexibility is not always better: different statistical regimes favor different classes. Our framework guides class selection using three measurable properties: how much improvement is possible beyond the best fixed action, whether there are enough samples to reliably detect gains from instance-level signals, and how much improvement a coarse partition can recover. We derive a finite-sample threshold for certifying selective prediction gains, with a matching-order lower bound, alongside bounds on residual improvement and partition gains. These results motivate a selection procedure that computes its diagnostics from training data alone. Across ten benchmark settings evaluated under strict nested cross-validation, the predicted class matches the empirical winner in nine. Replacing the judge on the same dataset shifts the diagnostics and reverses the winning class, as predicted. On TextVQA, an OCR experiment further separates the benefit of additional information from that of a deterministic gate, which outperforms the tested learners given the same information. Code is available at `https://github.com/Anonymous-Awesome-Submissions/Regime-Theory`.

1 INTRODUCTION

Consider a medical question-answering system in clinical use. When asked "what does this chest X-ray show?", the system has several options: answer directly, retrieve relevant prior cases first, defer to a stronger specialist model, or abstain. A system that always answers, even when uncertain, is dangerous. As large language and vision-language models move from benchmarks into deployment, *when* to answer is becoming as consequential as how. A confidently wrong diagnosis, a fluent but factually incorrect claim, or a hard query handled by a weak model when a stronger one was available all share a common cause: the system chose the wrong action for an uncertain input.

A common engineering response is to use a *router* that selects a model or processing path for each input (Chen et al., 2023; Ong et al., 2024; Dekoninck et al., 2024). Related approaches decide when to retrieve evidence (Lewis et al., 2020; Karpukhin et al., 2020; Izacard et al., 2023), defer to a stronger model (Mozannar & Sontag, 2020), or abstain (Geifman & El-Yaniv, 2017). These methods typically fix the form of the decision rule and optimize within it. Better scores or thresholds, however, do not establish whether the chosen structure is appropriate for the task and available data. They offer limited guidance on when a fixed action is sufficient, when decisions should be shared across groups, and when individual inputs warrant different decisions. We study this structural choice through a common abstraction: a *controller* is a policy that maps an input to an action.

With finite data, the most flexible structure is not always the best. Its value depends on whether the available signals distinguish which actions work better on different inputs, and whether those distinctions can be learned reliably. Prior work identifies conditions under which confidence-based deferral can be suboptimal (Jitkrittum et al., 2023), but does not determine which controller class

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

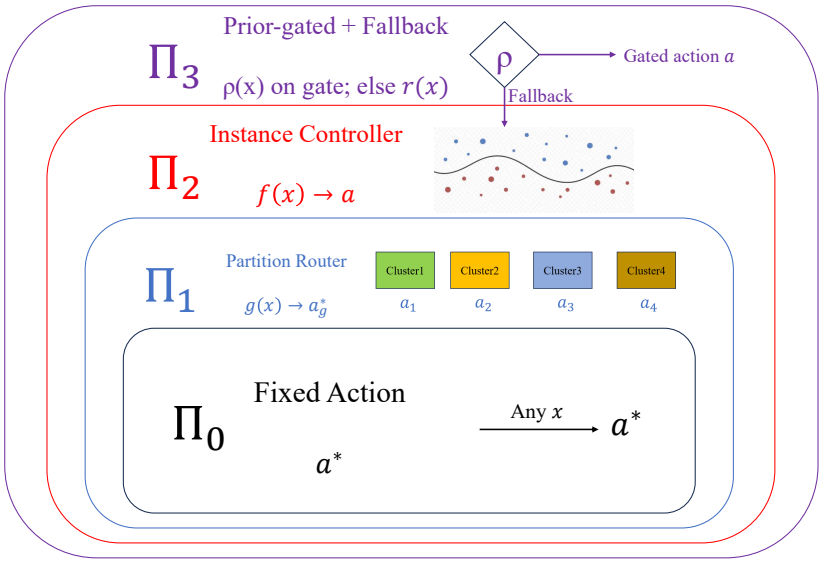

Figure 1: The nested lattice of policy classes. Π_0 contains fixed actions; Π_1 contains partition routers; Π_2 contains instance-level learned controllers; Π_3 pairs a deterministic prior gate with a fallback controller drawn from a lower class.

to choose at a given sample size. This motivates a framework that relates controller structure to the improvement available, the information in the features, and the amount of data.

We organize controllers into a strictly nested four-class lattice $\Pi_0 \subset \Pi_1 \subset \Pi_2 \subset \Pi_3$ (Figure 1): fixed actions, partition routers that split inputs into a few groups, instance-level learned controllers, and prior-gated controllers that use an external signal when it is confident and otherwise fall back to a lower-class controller. The statistical point is simple: more complex controllers require enough samples before their gains can be certified, but passing this viability test does not by itself make a class the right choice. At a given $(\mathcal{D}, n)$ we should use the class whose certified gain best exceeds its estimation cost, which can be a lower-complexity class even when a higher one is statistically viable. We derive a finite-sample viability threshold with a matching-order lower bound, together with bounds on the residual and partition gains (Section 2.2); the resulting diagnostics predict the empirical winner on nine of ten benchmark settings (Section 3).

2 Method

Section 2.1 defines the loss matrix and policy classes; Section 2.2 turns class choice into regime diagnostics and a nested-CV selector.

2.1 Problem Setup

We consider an input $x \sim \mathcal{D}$ and a finite action set $\mathcal{A} = \{a_{\text{direct}}, a_{\text{retrieve}}, a_{\text{defer}}, a_{\text{abstain}}\}$. Each action a has correctness $c(x,a) \in \{0,1\}$, semantic risk $h(x,a) \in [0,1]$, and operational cost $k(x,a) \geq 0$. The combined task loss is

$$\ell(x,a) = w_{\text{c}}(1 - c(x,a)) + w_{\text{h}} h(x,a) + w_{\text{k}} k(x,a) \in [0, L_{\max}],$$

with non-negative weights $(w_{\text{c}}, w_{\text{h}}, w_{\text{k}})$ fixed by the deployment. A controller $\pi : \mathcal{X} \to \mathcal{A}$ has population risk $R(\pi) := \mathbb{E}_{\mathcal{D}}[\ell(x, \pi(x))]$. Setting $w_{\text{h}} = w_{\text{k}} = 0$ recovers raw zero-one accuracy with a_{abstain} costing the same as a wrong direct prediction; classical reject-option theory (Chow, 2003; Franc et al., 2023; Pugnana & Ruggieri, 2023) is the special case in which a_{abstain} is assigned an intermediate cost $c_r \in (0,1)$ via the semantic-risk channel ($h(x, a_{\text{abstain}}) = c_r$, w_{h}=1).

Definition 1 (Policy classes)**.** Fix a candidate partition family $\mathfrak{G}$ and a deployed instance-level family $\mathcal{F}$ of measurable maps $\mathcal{X} \to \mathcal{A}$. Each class is a set of measurable maps $\mathcal{X} \to \mathcal{A}$:

$$\Pi_0 := \{\, \pi : \pi(x) \equiv a \text{ for some } a \in \mathcal{A} \,\},$$
$$\Pi_1 := \Pi_0 \cup \big\{\, \pi : \exists\, \mathcal{G} = \{G_g\}_{g=1}^{K} \in \mathfrak{G},\ (a_g)_{g \le K} \in \mathcal{A}^K,\ \pi|_{G_g} \equiv a_g \,\big\},$$
$$\Pi_2 := \Pi_1 \ \cup\ \mathcal{F},$$
$$\Pi_3 := \Pi_2 \cup \big\{\, \pi : \pi = \rho \text{ on } \{x : \text{the prior channel fires}\},\ \pi = r \text{ elsewhere},\ r \in \Pi_2 \,\big\},$$

where ρ is a deterministic rule on a channel not visible to r's feature inputs, and $\mathfrak{G}$ is the same candidate family that C_{Π_1} maximises over in Section 2.2. Nestedness $\Pi_0 \subset \Pi_1 \subset \Pi_2 \subset \Pi_3$ holds by construction.

On the capacity of Π_2. No result below invokes uniform convergence over $\mathcal{F}$, so no capacity notion—VC dimension, Rademacher complexity, or covering numbers—enters the analysis, and $\mathcal{F}$ is left unrestricted beyond measurability; what such a term would otherwise control is carried by scalar functionals of the deployed family and by a finite union bound (Appendix A).

For the set inclusion, view x as the full prediction-time observation; lower classes simply ignore side channels. The four rungs are a constant action, a partition, a per-sample rule, and a side-information gate. If no label-free prediction-time prior is available, the deployable lattice truncates to $\Pi_0 \cup \Pi_1 \cup \Pi_2$.

The classes are nested as sets of functions, so population-optimal risk is monotone: $R(\pi_0^\star) \ge R(\pi_1^\star) \ge R(\pi_2^\star) \ge R(\pi_3^\star)$. Estimation complexity goes the other way, and the question is which m minimizes the *expected* risk $\mathbb{E}_S[R(\widehat{\pi}_m(S))]$ at finite n, and whether the answer can be identified from data.

2.2 Finite-sample selection among policy classes

We isolate three bottlenecks: residual mass, Π_2 sample size, and Π_1 partition geometry. Three theorems characterize them, one lemma and one lower bound support the Π_2 analysis, and a final theorem bounds held-out class selection. Throughout, let $a^\star := \arg\min_{a \in \mathcal{A}} \mathbb{E}[\ell(x, a)]$ denote the best fixed action and $R := \{x : \exists a, \ell(x, a) < \ell(x, a^\star)\}$ the residual set.

We first bound how much any adaptive controller can improve over the best fixed action. If only a small fraction of inputs admit a better action, then the total gain is limited by that fraction times the largest possible per-input improvement.

Theorem 1 (Residual bound)**.** *For any controller $\pi : \mathcal{X} \to \mathcal{A}$,*

$$\mathbb{E}[\ell(x, a^\star) - \ell(x, \pi(x))] \ \le\ \mathbb{P}(R) \cdot \sup_{x \in R, a \in \mathcal{A}} (\ell(x, a^\star) - \ell(x, a)).$$

The proof (Appendix A) is a one-line conditioning argument. We next analyze a canonical selective subproblem inside Π_2 only to derive the finite-sample threshold; the evaluated Π_2 families in Section 3 remain multi-action.

A score f induces a rejector π_q: it plays a_{direct} on the top $1-q$ fraction of inputs, ranked by f, and a fallback action on the bottom q. Let L_r, L_w, L_a be the expected losses of a_{direct} on correct inputs, a_{direct} on wrong inputs, and the fallback. Define

$$\alpha_{\min} := \tfrac{L_a - L_r}{L_w - L_r}, \qquad \beta := \alpha_{\text{emp}} - \alpha_{\min},$$

where $\alpha(f)$ is the AUC of f against direct correctness and $\alpha_{\text{emp}} = \sup_f \alpha(f)$ is the AUC ceiling on the feature class. Thus $\alpha_{\min}$ is the break-even AUC for routing the bottom-q to a fallback, and $\beta = \alpha - \alpha_{\min}$ is the margin above break-even.

The asymptotic improvement condition is classical (Chow, 2003; Franc et al., 2023; Pugnana & Ruggieri, 2023): under the local tail-margin convention introduced in Lemma 1, π_q improves on a_{direct} asymptotically when $\beta > 0$, with reduction of order $q^\star \beta (L_w - L_r)$ at optimal $q^\star$ (the exact iff condition is on $\mu_w(q) - \alpha_{\min}$, for which β is the AUC-level scalar proxy).

Lemma 1 (Asymptotic signal condition, after (Chow, 2003; Franc et al., 2023)). *The exact improvement condition is on the bottom-q tail precision: π_q improves on a_{direct} asymptotically iff $\mu_w(q) > \alpha_{\min}$ for some q. Suppose $\mu_w(q) - \alpha_{\min}$ shares sign with the AUC margin $\beta = \alpha_{\text{emp}} - \alpha_{\min}$ at the operating q. We use this structural condition throughout as the local tail-margin convention. Under this convention, β is a sufficient scalar proxy: $\beta \leq 0$ rules out asymptotic improvement, and $\beta > 0$ admits scale $q^\star \beta(L_w - L_r)$ at optimal coverage $q^\star$.*

The new finite-sample result quantifies how large n must be for the population gain to be visible above estimator variance.

Theorem 2 (Bernstein-tight viability threshold). *Let $\ell \in [0, L_{\max}]$, $\alpha \geq 1/2$, and $\beta = \alpha - \alpha_{\min} > 0$. Assume the local tail-margin condition $\mu_w(q) - \alpha_{\min} \geq \beta$, where $\mu_w(q) := \mathbb{P}(c_{\text{direct}}{=}0 \mid f(x) \in \text{bottom-}q)$. Then to leading order in n,*

$$n \;\geq\; n_{\min}(\alpha, \beta, q, \delta) \;:=\; \frac{2\alpha(1-\alpha)\log(2/\delta)}{q\beta^2} \tag{1}$$

is sufficient for the empirical CV loss difference $\widehat{\Delta}(q)$ to have the same sign as the population loss difference with probability at least $1 - \delta$. When n is materially below (1)*, the bound no longer certifies sign correctness; we call this regime* variance-bounded.

The threshold is computable from $(\alpha, \beta, q, \delta)$ alone (Figure 2). Equation (1) is the leading-order form; solving the exact condition raises $n_{\min}$ by a factor of 1.2–2.4 on our benchmarks without moving any verdict (Appendix A.3). Theorem 2 is the finite-sample counterpart of Jitkrittum et al. (2023): even with positive asymptotic margin, the sign of a controller's empirical gain can be uncertified when n is too small. The proof (Appendix A.3) applies Bernstein's inequality to the bottom-q precision estimator. The next proposition shows that the same dependence on α, β, q, and δ is necessary up to constants.

Proposition 1 (Information-theoretic lower bound). *In the setting of Theorem 2, suppose the bottom-q decision reduces locally to testing $H_0 : \mu_w = \alpha_{\min}$ against $H_1 : \mu_w = \alpha_{\min} + \beta$, with $0 < \beta \leq \frac{1}{2}\min\{\alpha_{\min}, 1 - \alpha_{\min}\}$. Any test that determines the sign of the population loss difference $\Delta(q)$ with both type-I and type-II error at most $\delta \in (0, \frac{1}{8})$ from n i.i.d. samples must satisfy*

$$n \;\geq\; \frac{c\,\alpha(1-\alpha)\,\log(1/\delta)}{q\,\beta^2} \tag{2}$$

for a universal constant $c > 0$. Since the sufficient condition (1) *has the same functional form $n_{\min} = 2\alpha(1-\alpha)\log(2/\delta)/(q\beta^2)$, the viability threshold has the optimal high-confidence order under this reduction: $n_{\min} \asymp \alpha(1-\alpha)\log(1/\delta)/(q\beta^2)$, up to universal constants and the local replacement of $\alpha_{\min}$ by $\alpha = \alpha_{\min} + \beta$.*

The proof (Appendix A.4) is Bretagnolle–Huber on the local Bernoulli alternative (Tsybakov, 2009).

A coarser controller can win where Π_2 is variance-bounded because it asks less of the data: a partition router needs only a clear best action per group, so its sample complexity is governed by cluster mass and gap rather than by the per-sample margin that controls Π_2. Let $\mathcal{G} = \{G_g\}_{g=1}^K$ be a measurable partition with $p_g := \mathbb{P}(x \in G_g)$, cell-best actions $a_g^\star := \arg\min_a \mathbb{E}[\ell(x, a) \mid x \in G_g]$, and discriminability gaps $\gamma_g := \mathbb{E}[\ell(x, a^\star) - \ell(x, a_g^\star) \mid x \in G_g] \geq 0$.

Theorem 3 (Distribution-dependent partition lower bound). *For any measurable partition $\mathcal{G}$ with cluster probabilities p_g and discriminability gaps γ_g,*

$$\Delta_{\text{partition}}(\mathcal{G}) \;=\; \sum_{g=1}^{K} p_g \gamma_g \;\geq\; \max_g\, p_g \gamma_g. \tag{3}$$

The sample complexity depends on p_g and γ_g but not on the instance-level margin β.

A partition router can therefore succeed precisely where an instance-level controller is variance-bounded. The $\max_g p_g\gamma_g$ form is also diagnostic: when the empirical reduction is close to it, the partition win is essentially a one-cluster phenomenon. Proof is in Appendix A; the bound is agnostic to how the partition is chosen, and an adaptive partition such as KMeans can be tighter; we establish that empirically rather than in the bound (Appendix C.2).

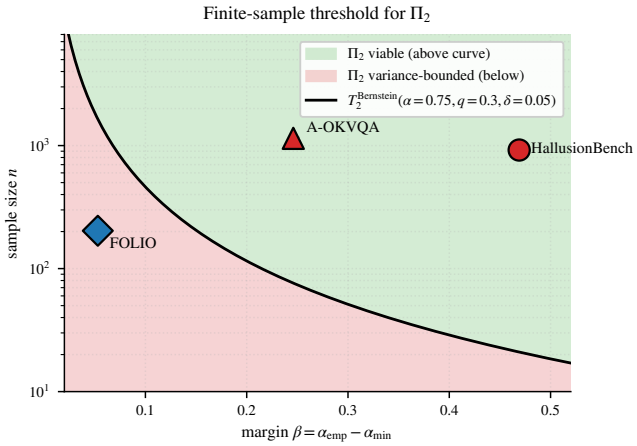

Figure 2: The viability threshold from (1) at α=0.75, q=0.3, δ=0.05. Green: Π_2-viable. Red: variance-bounded. FOLIO is the one core benchmark far below the Bernstein threshold (n=203 vs. $n_{\min}$=1898) and lands deep in the variance-bounded region; its best empirical Π_2 controller fails to improve on Π_0, the uncertified-sign regime of Theorem 4 (ii). HallusionBench (n=920 $\gg n_{\min}$=23) and A-OKVQA (n=1145 $\gg n_{\min}$=45) sit above their respective thresholds; whether their empirical margins separate is a further question, taken up in Table 1. SMS-Spam is governed instead by the residual bound of Theorem 1: its $\beta \leq 0$ violates the positive-margin condition of Theorem 2, so it lies outside the theorem's viability-threshold regime.

For comparing supportable adaptive classes, write

$$C_{\Pi_1}(\mathcal{G}) := \sum_g p_g \gamma_g, \qquad C_{\Pi_1} := \max_{\mathcal{G}\in\mathfrak{G}} C_{\Pi_1}(\mathcal{G}),$$

for a candidate partition family $\mathfrak{G}$, and

$$C_{\Pi_2} := q^\star \beta (L_w - L_r)$$

for the AUC-level Π_2 ceiling. The residual product, the Π_2 viability test ($n \gtrless n_{\min}$), and the two ceilings C_{Π_1}, C_{Π_2} combine into the operational result of this paper. Theorems 1–3 each characterise one bottleneck; the following theorem combines them into a class choice, in the form the prospective selector of Section 3.1 computes:

Theorem 4 (Regime decomposition)**.** *For $(\mathcal{D}, n)$, (i) every adaptive class is capped by the residual product in Theorem 1; when this product is small, no adaptive class can improve by more than residual scale. (ii) If $\beta > 0$ but $n < n_{\min}$, Π_2's sign is uncertified, while Π_1 can strictly improve on Π_0 whenever some candidate partition has $C_{\Pi_1}(\mathcal{G}) > 0$. (iii) If $n \geq n_{\min}$, the population-optimal winner between Π_1 and Π_2 is determined by comparison between the ceilings C_{Π_1} and C_{Π_2}; the empirical winner additionally depends on how tightly the deployed candidate families realize each ceiling at finite n.*

Section 3 uses strict nested CV over candidate classes, and Theorem 5 bounds the held-out selection cost. The bound is stated for the inner-CV estimand because validation-based selection can itself overfit (Cawley & Talbot, 2010).

Theorem 5 (Held-out selection bound)**.** *Let $\mathcal{M} \subseteq \{0, 1, 2, 3\}$ be the candidate classes ($|\mathcal{M}| = 3$ when no deployable external channel is available). Let $\widehat{R}_{\mathrm{CV}}(m)$ be the inner κ-fold CV estimate and $\bar{R}_\kappa(m)$ the expected risk of an inner-trained estimator. With $\widehat{m} := \arg\min_{m\in\mathcal{M}} \widehat{R}_{\mathrm{CV}}(m)$ and $\ell \in [0, L_{\max}]$,*

$$\mathbb{E}[\bar{R}_\kappa(\widehat{m})] \;\leq\; \min_{m\in\mathcal{M}} \bar{R}_\kappa(m) + 4L_{\max}\sqrt{\log(2|\mathcal{M}|)/n_{\mathrm{in}}}, \tag{4}$$

where $n_{\mathrm{in}} = n_{\mathrm{out}}/\kappa$. If each estimator is $\varepsilon_{\mathrm{stab}}$-stable when refit from inner-train to outer-train, the bound transfers to the refit estimator with an added $2\varepsilon_{\mathrm{stab}}$ term.

The proof (Appendix A.6) is Hoeffding plus a union bound over $\mathcal{M}$. The $\min_m$ on the right absorbs the class-specific estimation errors controlled by Theorems 2–3: zero for $m{=}0$, partition complexity for $m{=}1$, and the Bernstein threshold for $m{=}2$.

3 EXPERIMENTS

Experimental protocol and data. We instantiate the lattice as a controller pool under strict nested 5-fold-by-5-seed CV: family selection on outer-train via inner CV, refit on outer-train, single evaluation on outer-test. Hyperparameters are fixed in advance, not tuned per benchmark. Reported $\pm$ values are 5-seed standard deviations; gaps in seed-sd units reflect algorithmic stability rather than dataset-resampling error. The family pool, feature blocks, loss-matrix construction and full protocol are in Appendices B–B.4; cost-sensitive L2D adaptations (Mozannar & Sontag, 2020; Narasimhan et al., 2022) are in Appendix E.1.

The four core benchmarks are SMS-Spam (Almeida et al., 2011) (saturated text, residual-bounded), HallusionBench (Guan et al., 2024) (visual hallucination, large-β), A-OKVQA (Schwenk et al., 2022) (near-tie Π_1/Π_2), and FOLIO (Han et al., 2024) (small β, $n{<}n_{\min}^{\mathrm{Bern}}$, variance-bounded), each on its full native held-out split. Each regime tag follows from $(\alpha, \beta, n, \{p_g\gamma_g\})$ computed on the held-out split, independent of any per-class CV outcome. Actions come from Qwen2.5-VL-3B/7B-Instruct; per-action semantic risk on HallusionBench and A-OKVQA is labelled by an InternVL2.5-8B (Chen et al., 2024b) judge *drawn from a different VLM family than the upstream policies*, and that judge never enters the controller's feature block. None of the four supplies a label-free prediction-time prior channel, so the deployable lattice truncates to $\Pi_0 \cup \Pi_1 \cup \Pi_2$; we use TextVQA (Singh et al., 2019) for a separate Π_3 test.

Main result. Table 7 gives the deployable per-class winners and Figure 7 the diagnostics that predict them; the pattern aligns with Theorem 4. ***SMS-Spam*** (residual-bounded, $\mathbb{P}(R){=}0.009$): every Π_1 family collapses to direct in CV and the best Π_2 family is $+0.0002$ worse, so Π_0 is the winner, by a margin the seeds cannot resolve (Appendix C.1). ***HallusionBench*** (Bernstein-viable, $n{=}920 \gg n_{\min}^{\mathrm{Bern}}{=}23$): Π_2 (HGBC-`md3`, -0.103, 26 seed-sd) edges Π_1 (KMeans-$K{=}8$, -0.091) by ≈ 3 seed-sd; the instance ceiling exceeds the partition ceiling (Thm 4 (iii)). ***A-OKVQA*** (Π_1/Π_2 boundary): Π_2 (Selective-$C{=}0.3$, -0.033) edges Π_1 (KMeans-$K{=}4$, -0.024) by ≈ 1 seed-sd, a gap the seeds do not resolve. ***FOLIO*** (variance-bounded, $n{=}203 \ll n_{\min}^{\mathrm{Bern}}{=}1898$; Thm 4 (ii); text-specialised-backbone replication in Appendix E.8): Π_1 (KMeans-$K{=}6$, -0.033, 4.5 seed-sd) wins; the best Π_2 controller increases loss relative to Π_0 by $+0.0027$, and both cost-sensitive L2D adaptations (Mozannar & Sontag, 2020; Narasimhan et al., 2022), instance-level and so in Π_2, trail our Π_1 (Appendix E.1). The A-OKVQA Π_3 rationale experiment is excluded from the deployable comparison because rationales are answer-derived (Appendix C.4). The classification is unchanged at $q \in \{0.2, 0.3, 0.4\}$.

TextVQA-OCR: the Π_3 channel condition. Π_3 needs a label-free prediction-time prior channel, and TextVQA supplies one: OCR tokens are image-derived and observable before answering. Routing on the official validation split ($n{=}5000$), the OCR gate reduces loss by 0.0134 over the best lower class, which sees only non-OCR features—6.4 joint seed-sd (Table 11). That gain has two sources, and both are resolved. Handing the OCR block to every class as an ordinary feature lifts the best lower class by 0.0073, which is what the channel is worth; the gate keeps a further 0.0060 over that, 2.6 joint seed-sd, which is what the gating form is worth (Appendix E.7). Giving the same channel to a learner recovers less of it at every capacity we tried. The criterion this yields is checkable before any controller is fitted: acquire a label-free channel when its incremental AUC over the deployed feature block is material, and fold it into Π_2 as an ordinary feature when it is not. A-OKVQA supplies the negative end—none of three screened channels qualifies, and gating on the strongest of them loses to the Π_2 winner (Appendix C.4). This is why the four core benchmarks truncate to $\Pi_0 \cup \Pi_1 \cup \Pi_2$.

3.1 PROSPECTIVE CLASS SELECTION AND GENERALISATION

The agreement reported above is read off diagnostics computed on outer-test concatenations, so it certifies the theory but not a deployable procedure. We therefore built the prospective form: inside each outer fold the Theorem 4 diagnostics are recomputed from outer-train alone, before any family

Benchmark	n	predicted	empirical	margin	selector class	selector Δ
RouterBench	36,497	Π_2	Π_2 ✓	23.4 sd	Π_2 25/25	-0.133
POPE-ext	9000	Π_2	Π_2 ✓	8.9 sd	Π_2 25/25	-0.016
DBpedia-14	5000	Π_1	Π_1 ✓	19.3 sd	—	—
FEVER	1415	Π_1	Π_1 ✓	0.5 sd	Π_1 25/25	-0.181
A-OKVQA	1145	Π_2	Π_2 ✓	1.1 sd	Π_2 25/25	-0.031
SMS-Spam	1114	Π_0	Π_0 ✓	—	Π_0 25/25	0.000
HallusionBench	920	Π_2	Π_2 ✓	2.7 sd	Π_2 25/25	-0.187
HallusionBench (2nd judge)	920	Π_1	Π_1 ✓	3.6 sd	Π_1 25/25	-0.033
SST-2	872	Π_1	Π_2 ✗	2.4 sd	—	—
FOLIO	203	Π_1	Π_1 ✓	4.5 sd	Π_1 25/25	-0.023

Table 1: All ten benchmarks, ordered by n. "Margin" is the gap between the empirical winner and the runner-up class in joint seed-sd units; SMS-Spam's classes are numerically tied so no margin is defined. The selector columns report the prospective form—diagnostics recomputed inside each outer fold from outer-train alone—on the eight benchmarks for which we ran it, giving the class it chose on all 25 outer folds and its loss reduction against always-direct. The prediction holds on nine of ten, and where the selector was run it reproduced the predicted class on every fold. Two matches, FEVER at 0.5 sd and A-OKVQA at 1.1 sd, are not resolved above seed noise: there the prediction agrees with the empirical argmin but the two classes are statistically tied. The diagnostics behind every row are plotted in Figure 7.

is fitted; the candidate pool is restricted to the class they name, and inner CV then chooses a family within that class. We also ran six further datasets through the identical pipeline, spanning two orders of magnitude in n and including the routing literature's own benchmark (RouterBench, all 86 eval sets pooled). Table 1 reports both.

The pruning the diagnostic performs is benchmark-dependent and visible per fold: on SMS-Spam the residual test admits Π_0 alone on all 25 folds; on HallusionBench and A-OKVQA it admits all three classes, so the restriction is vacuous and the selector reduces to unrestricted inner CV; on FOLIO it excludes Π_2 on all 25 folds, on 20 through $n < n_{\min}$ and on 5 through $\beta \leq 0$. FOLIO is where the restriction does the work, lifting the auto-pick from $\Delta = -0.003$ to -0.0227 against -0.0325 for the best Π_1 family identified post-hoc; the remaining gap is family choice inside Π_1, not class choice (Appendix E.3).

What the added benchmarks establish. RouterBench puts the diagnostics on the routing literature's own evaluation set two orders of magnitude above the core four, and the Π_2 prediction is confirmed at 23.4 seed-sd. DBpedia-14 is the sharpest quantitative check in the collection: at n=5000 against $n^{\text{Bern}}_{\min}$=43,732 it sits deep in the variance-bounded branch, and the predicted ceiling and the realised gain coincide to three digits (C_{Π_1}=0.0097 against a Π_1 reduction of 0.0097 over the fixed action). FEVER is the only benchmark that lands in branch (iii) with $C_{\Pi_1} > C_{\Pi_2}$ while sitting far above the threshold, the configuration under which a partition router should win on ceiling grounds rather than on small-n grounds; its Π_1–Π_2 gap is not resolved above seed noise, and we return to what that limits in Section 5.

A judge ablation as a controlled regime flip. Re-judging the same 920 HallusionBench items with Phi-3.5-vision-instruct, a third model family disjoint from both the upstream policies and the incumbent judge, changes the loss matrix and leaves the items, the policies, the features and the protocol untouched. Every diagnostic moves: β falls from 0.469 to 0.036, $n^{\text{Bern}}_{\min}$ rises from 23 to 3726 so that n=920 crosses from above the threshold to well below it, and the ceiling comparison inverts. The empirical winner inverts with them, from Π_2 at 2.7 seed-sd over Π_1 to Π_1 at 3.6 seed-sd over Π_2.

SST-2: a miss, and what it localises. On SST-2 the rule predicts Π_1 and Π_2 wins, by 0.0069 or 2.4 joint seed-sd—outside seed noise, so this is a failure of the prediction and not of resolution. Here $\beta = -0.136 < 0$, so Lemma 1 rules out an instance-level *rejector* and the derived ceiling $C_{\Pi_2} = -0.041$ is negative; but β is computed for the binary reject-versus-answer subproblem, whereas the deployed Π_2 controller has four actions and can route rather than only reject. On SST-2

that gap is large enough to flip the verdict: an instance-level controller with no usable rejection signal still realises 0.0134 of gain over the best fixed action by redistributing inputs across the action set. The error is one-sided—β understates multi-action gain and so favours the coarser class—and a practitioner can see the failure coming from its profile: small β, a rich action set, and a residual set large enough ($\mathbb{P}(R) = 0.089$) to leave room for redistribution.

Routing baselines. Three canonical routing designs (Chen et al., 2023; Ong et al., 2024; Jitkrittum et al., 2023) run through the identical protocol. We place all three in Π_2 because each scores an input individually and thresholds that score, which is the defining form of Definition 1's instance-level class; the outcome then follows the regimes: where the diagnostics select Π_2 they match us to within 0.023; where a coarser class is selected they trail the Π_1 router on FOLIO by at least 0.032, and two of three increase loss over always-direct on residual-bounded SMS-Spam (Appendix E.5).

Additional checks. Theorem 3's bound is empirically tight on HallusionBench: $\sum_g p_g\gamma_g{=}0.047$ at $K{=}4$ against a realised KM-$K{=}4$ reduction of 0.048 ± 0.005 (Appendix C.2). Three controlled synthetics sweep n across the viability threshold, vary instance signal at fixed partition gap, and vary channel informativeness (Appendices D.1–D.3); CART and loss-weight ablations are in Appendix E.2.

4 Related Work

Selective classification, learning to defer, and cascade routing. A long line of work fixes a single controller family and refines the score, surrogate, or calibrator within it—from classical reject-option theory (Chow, 2003; El-Yaniv et al., 2010; Geifman & El-Yaniv, 2017; Franc et al., 2023; Pugnana & Ruggieri, 2023; Traub et al., 2024) to cost-sensitive deferral (Mozannar & Sontag, 2020; Narasimhan et al., 2022; Verma et al., 2023; Mao et al., 2023). Cascade routing (Chen et al., 2023; Ding et al., 2024; Ong et al., 2024; Chen et al., 2024a; Dekoninck et al., 2024) extends this to a cost-ordered model sequence but keeps the family fixed, Goren et al. (2024) nest abstention in the *output* space, and conformal prediction (Angelopoulos & Bates, 2023; Bates et al., 2021) calibrates a threshold inside a fixed selective family. We ask which family the available n can support, and our held-out selection bound (Theorem 5) carries classical nested-CV concerns (Cawley & Talbot, 2010) from hyperparameter selection to class selection.

Gate-and-residual hybrids and hallucination impossibility. Combining a deterministic gate with a learned fallback recurs across selective generation (Chen et al., 2023; Dekoninck et al., 2024; Jeong et al., 2024; Asai et al., 2023). Π_3 singles out the subclass with a *deterministic* rule on a channel disjoint from the fallback's feature block, and gives the regime-theoretic conditions under which this restriction strictly dominates a learned Π_2—and when the Π_1 fallback wins instead. Where instance-level signal is structurally limited (Kalai & Vempala, 2024; Karbasi et al., 2025), Theorem 2 supplies the finite-sample counterpart and Theorem 3 identifies when Π_1 can still recover loss; mitigations elsewhere in that literature act on decoding (Leng et al., 2024), features (Farquhar et al., 2024), or the action set (Srinivasan et al., 2024; Whitehead et al., 2022).

5 Limitations

Semantic-risk labels on HallusionBench and A-OKVQA come from a single judge, and replacing it with one from a third model family moves HallusionBench's regime tag and its empirical winner together (Section 3.1). Where the two ceilings are commensurate, Theorem 4 (iii)'s prediction is population-optimal but mediated at finite n by how well the deployed families realize them (Appendix D.2).

Three further limits deserve separate statement. First, the class gate β is derived from a binary reject-versus-answer subproblem while the deployed controllers are four-action; the SST-2 miss of Section 3.1 is where that gap decides the verdict, and Appendix E.6 measures it directly. Second, branch (iii) on the $C_{\Pi_1} > C_{\Pi_2}$ side rests on the direction of the ceiling comparison and not yet on a resolved empirical separation: FEVER is our only benchmark in that configuration, and its Π_1–Π_2 gap sits inside seed noise. Third, the prospective selector is evaluated on the same benchmark pool from which the diagnostics were designed, so its per-fold agreement is internal consistency

of the rule and not an out-of-distribution guarantee. The Π_3 rung additionally requires a label-free prediction-time channel, which only two of our benchmarks supply.

6 Conclusion and Discussion

We presented a framework for selecting controller classes for language and vision-language models. Across fixed actions, partition routers, instance-level controllers, and prior-gated controllers, greater flexibility does not consistently produce better decisions with finite data. Our analysis connects class selection to the improvement available beyond a fixed action, the reliability of instance-level signals, and the gains recoverable through partitions. Theoretical bounds turn these considerations into diagnostics computable from training data alone, and the experiments show that they anticipate the preferred class. The one miss, on SST-2, traces to deriving the instance-level test from a binary reject-versus-answer subproblem; a test built on the full action set is the natural next step.

The preferred controller class is a property of the complete decision problem. It depends on the available actions, the costs assigned to their outcomes, the information observable at prediction time, and the sample budget. Changing these conditions can change which distinctions between inputs are worth learning. Controller selection should therefore be revisited as a deployment evolves. This perspective gives routing, deferral, and abstention a common basis for comparison: the improvement a decision structure can recover and the evidence required to recover it reliably.

Reproducibility Statement

All theoretical claims are stated with their assumptions in Section 2 and proved in full in Appendix A, which contains complete proofs of Theorems 1–5, Lemma 1, and Proposition 1. The empirical protocol is specified in Appendix B: Appendix B.4 gives the strict nested 5-fold-by-5-seed cross-validation procedure, seeds, and the single-evaluation rule that prevents test reuse; the feature blocks, loss-matrix construction, and the full family pool with all fixed hyperparameters are given in the preceding subsections. Per-benchmark data preparation, including the upstream policy backbones, the judge model, and the exact held-out splits and row counts, is documented in Appendix C. Baseline adaptations, ablations, and the automated class-selection diagnostics are in Appendix E.2 and Appendix E.3. Anonymized code reproducing every table and figure is available at the repository linked in the abstract.

Ethics Statement

This work uses only publicly released benchmarks (SMS-Spam, HallusionBench, A-OKVQA, FOLIO, TextVQA, RouterBench, POPE-ext, DBpedia-14, FEVER, and SST-2) under their original licenses, and involves no human subjects, no new data collection, and no release of new data. The controllers studied here decide whether a deployed model should answer, retrieve, defer, or abstain; the motivating applications include high-stakes settings such as clinical question answering. We note two ethical considerations. First, the loss matrices used here encode operating costs that are application-specific: a class-selection result obtained under one cost setting should not be transferred to a deployment with materially different error costs without re-estimating the diagnostics, and Appendix E.2 reports the sensitivity of our conclusions to those weights. Second, an abstention or deferral policy shifts rather than removes risk, and its downstream effect depends on what the fallback path actually is; we make no claim that the controllers evaluated here are safe to deploy in any specific clinical or otherwise high-stakes system without domain-specific validation.

AI Use Statement

In this work, we used generative AI tools for code creation and editing, and for editing the paper for clarity; both fall under ICLR's recommended-disclosure categories. We have not used generative AI tools for the tasks that require disclosure, and the remaining required-disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Tiago A Almeida, José María G Hidalgo, and Akebo Yamakami. Contributions to the study of sms spam filtering: new collection and results. In *Proceedings of the 11th ACM symposium on Document engineering*, pp. 259–262, 2011.

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

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2023.

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

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 02 2013. ISBN 9780199535255. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

Shuhao Chen, Weisen Jiang, Baijiong Lin, James Kwok, and Yu Zhang. Routerdc: Query-based router by dual contrastive learning for assembling large language models. *Advances in Neural Information Processing Systems*, 37:66305–66328, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

C Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 2003.

Jasper Dekoninck, Maximilian Baader, and Martin Vechev. A unified approach to routing and cascading for llms. *arXiv preprint arXiv:2410.10347*, 2024.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Ruhle, Laks VS Lakshmanan, and Ahmed Hassan Awadallah. Hybrid llm: Cost-efficient and quality-aware query routing. *arXiv preprint arXiv:2404.14618*, 2024.

Chris Drummond and Robert C Holte. Cost curves: An improved method for visualizing classifier performance. *Machine learning*, 65(1):95–130, 2006.

Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Vojtech Franc, Daniel Prusa, and Vaclav Voracek. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Shani Goren, Ido Galil, and Ran El-Yaniv. Hierarchical selective classification. *Advances in Neural Information Processing Systems*, 37:111047–111073, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, et al. Folio: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22017–22031, 2024.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43, 2023.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, 2024.

Wittawat Jitkrittum, Neha Gupta, Aditya Krishna Menon, Harikrishna Narasimhan, Ankit Singh Rawat, and Sanjiv Kumar. When does confidence-based cascade deferral suffice?, 2024. *URL https://arxiv. org/abs/2307.02764*, 2023.

Adam Tauman Kalai and Santosh S Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 160–171, 2024.

Amin Karbasi, Omar Montasser, John Sous, and Grigoris Velegkas. (im) possibility of automated hallucination detection in large language models. *arXiv preprint arXiv:2504.17004*, 2025.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6769–6781, 2020.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. *Advances in neural information processing systems*, 36:3578–3606, 2023.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, pp. 7076–7087. PMLR, 2020.

Harikrishna Narasimhan, Wittawat Jitkrittum, Aditya K Menon, Ankit Rawat, and Sanjiv Kumar. Post-hoc estimators for learning to defer to an expert. *Advances in Neural Information Processing Systems*, 35:29292–29304, 2022.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms with preference data. *arXiv preprint arXiv:2406.18665*, 2024.

Andrea Pugnana and Salvatore Ruggieri. Auc-based selective classification. In *International conference on artificial intelligence and statistics*, pp. 2494–2514. PMLR, 2023.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pp. 146–162. Springer, 2022.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. Selective "selective prediction": Reducing unnecessary abstention in vision-language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12935–12948, 2024.

Jeremias Traub, Till J Bungert, Carsten T Lüth, Michael Baumgartner, Klaus H Maier-Hein, Lena Maier-Hein, and Paul F Jäger. Overcoming common flaws in the evaluation of selective classification systems. *Advances in Neural Information Processing Systems*, 37:2323–2347, 2024.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Rajeev Verma, Daniel Barrejón, and Eric Nalisnick. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 11415–11434. PMLR, 2023.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2022.

A Proofs

A.1 Theorem 1 (residual bound)

Let $g(x) := \ell(x, a^\star) - \ell(x, \pi(x))$. On R^c, by the definition of the residual set, $\ell(x, a^\star) \le \ell(x, a)$ for every $a \in \mathcal{A}$, so $g(x) \le 0$ on R^c. Therefore

$$\mathbb{E}[g(x)] = \mathbb{E}[g(x)\mathbf{1}_R] + \mathbb{E}[g(x)\mathbf{1}_{R^c}] \le \mathbb{E}[g(x)\mathbf{1}_R] \le \mathbb{P}(R) \sup_{x \in R, a \in \mathcal{A}} \big(\ell(x, a^\star) - \ell(x, a)\big),$$

which is the stated bound. □

A.2 Lemma 1 (asymptotic signal condition)

The Chow-style derivation in (Chow, 2003) and its selective-risk extensions (Franc et al., 2023; Pugnana & Ruggieri, 2023) give that the population loss difference of the rejector at coverage q induced by f, relative to the fixed direct policy, is

$$\Delta L(q) = -q\,(L_w - L_r)\,(\mu_w(q) - \alpha_{\min}),$$

where $\mu_w(q)$ is the population precision of the rejected set. Thus the exact improvement condition is $\mu_w(q) > \alpha_{\min}$. The AUC margin $\beta = \alpha_{\text{emp}} - \alpha_{\min}$ used in the main text is a scalar proxy for this working-point margin, not an identity between AUC and tail precision. Under the local tail-margin convention stated before Theorem 2, the certified loss-reduction scale is $q^\star \beta (L_w - L_r)$ after optimizing over $q \in (0, 1]$. □

A.3 Theorem 2 (Bernstein-tight viability threshold)

Notation. Let

$$m := \lfloor nq \rfloor, \qquad d := L_w - L_r > 0, \qquad Y_i := \mathbf{1}\{c_{\text{direct},i} = 0\}, \qquad \widehat{\mu}_w(q) := \frac{1}{m} \sum_{i=1}^{m} Y_i,$$

where the m observations are those the fitted score f places in the bottom-q set. Write $t := \log(2/\delta)$ and $\alpha := \alpha_{\min} + \beta \ge \frac{1}{2}$.

Step 1 (conditioning). Condition throughout on the σ-field $\mathcal{C}$ generated by the fitted score f, the resulting bottom-q index set, and the constants L_w, L_r, L_a. Under the nested-CV protocol of Appendix B.4, f is fitted on a split disjoint from the m evaluation points, so the selected set is determined independently of the labels used to evaluate it. Conditionally on $\mathcal{C}$, the Y_i are therefore independent Bernoulli variables with means

$$p_i := \mathbb{E}[Y_i \mid \mathcal{C}], \qquad \mu_w(q) := \frac{1}{m}\sum_{i=1}^{m} p_i.$$

Step 2 (population sign). Define

$$\Delta(q) := -qd\big(\mu_w(q) - \alpha_{\min}\big), \qquad \widehat{\Delta}(q) := -qd\big(\widehat{\mu}_w(q) - \alpha_{\min}\big).$$

The local tail-margin hypothesis of the theorem statement gives

$$\mu_w(q) - \alpha_{\min} \;\geq\; \beta \;>\; 0 \qquad \Longrightarrow \qquad \Delta(q) \;\leq\; -qd\beta \;<\; 0. \tag{A}$$

Step 3 (variance bound, and where $\alpha \geq 1/2$ is used). The conditional variance of the sample mean is $m^{-2}\sum_i p_i(1-p_i)$, so the relevant variance proxy is $m^{-1}\sum_i p_i(1-p_i)$. Since $u \mapsto u(1-u)$ is concave, Jensen gives the first inequality below; since the same map is decreasing on $[\frac{1}{2}, 1]$ and the local tail-margin hypothesis places $\mu_w(q) \geq \alpha \geq \frac{1}{2}$, it gives the second:

$$\frac{1}{m}\sum_{i=1}^{m} p_i(1-p_i) \;\leq\; \mu_w(q)\big(1-\mu_w(q)\big) \;\leq\; \alpha(1-\alpha) \;=:\; v. \tag{B}$$

Without $\alpha \geq \frac{1}{2}$ the second inequality can fail and v must be taken as $\frac{1}{4}$.

Step 4 (concentration). Set

$$r_m(\delta) := \sqrt{\frac{2vt}{m}} + \frac{2t}{3m}, \qquad \mathcal{E}_\delta := \Big\{\big|\widehat{\mu}_w(q) - \mu_w(q)\big| \leq r_m(\delta)\Big\}.$$

Bernstein's inequality (Boucheron et al., 2013; Vershynin, 2018) applied conditionally on $\mathcal{C}$, with variance proxy v from (B) and range 1, gives two-sided failure probability $2e^{-t} = \delta$:

$$\mathbb{P}\big[\mathcal{E}_\delta \mid \mathcal{C}\big] \;\geq\; 1 - \delta. \tag{C}$$

Step 5 (sign preservation). On $\mathcal{E}_\delta$,

$$\widehat{\mu}_w(q) - \alpha_{\min} \;\geq\; \big(\mu_w(q) - \alpha_{\min}\big) - r_m(\delta) \;\geq\; \beta - r_m(\delta),$$

using (A) for the second inequality. Hence if

$$r_m(\delta) \;<\; \beta \tag{D}$$

then $\widehat{\Delta}(q) < 0$. Combining (A), (C) and (D) and taking expectations over $\mathcal{C}$,

$$\mathbb{P}\big[\operatorname{sign}\widehat{\Delta}(q) = \operatorname{sign}\Delta(q)\big] \;\geq\; 1 - \delta. \tag{E}$$

Step 6 (solving (D) exactly). Substituting $s := \sqrt{m}$, condition (D) reads $\beta s^2 - \sqrt{2vt}\, s - \frac{2t}{3} > 0$, a quadratic in s with a single positive root. Therefore (D) holds iff

$$m \;>\; m_{\text{exact}} \;:=\; \left(\frac{\sqrt{2vt} + \sqrt{2vt + \frac{8}{3}\beta t}}{2\beta}\right)^2, \tag{F}$$

and since $m = \lfloor nq \rfloor$ it suffices that

$$n \;>\; \frac{m_{\text{exact}} + 1}{q}. \tag{G}$$

Conditions (F)–(G) are the exact sufficient form; they involve no approximation and are what the per-benchmark thresholds in Table 2 are solved from.

Benchmark	α	β	leading-order $n_{\min}$	exact $n_{\min}$ from (G)	n
HallusionBench	0.722	0.469	23	55	920
A-OKVQA	0.874	0.246	45	104	1145
FOLIO	0.686	0.053	1898	2201	203

Table 2: Leading-order (1) against the exact sufficient condition (G), at q=0.3, δ=0.05. Dropping the $1/n$ term costs a factor of 1.16–2.39. No verdict moves: against the exact thresholds HallusionBench sits at 16.7×, A-OKVQA at 11.0×, and FOLIO at 0.09×, the last slightly deeper into the variance-bounded regime than the leading-order figure indicated.

Step 7 (leading order). Expanding (F) as $\beta \to 0$ with v, t fixed, writing $\sqrt{2vt + \frac{8}{3}\beta t} = \sqrt{2vt}\,(1 + \frac{2\beta}{3v} - \frac{2\beta^2}{9v^2} + O(\beta^3))$,

$$m_{\text{exact}} = \frac{2vt}{\beta^2} + \frac{4t}{3\beta} - \frac{2t}{9v} + O(\beta),$$

so that, with $v = \alpha(1-\alpha)$ and $t = \log(2/\delta)$, the leading term of (G) is

$$n \;\gtrsim\; \frac{2\alpha(1-\alpha)\log(2/\delta)}{q\beta^2},$$

which is (1). Note that the $O(1)$ term is $-2t/(9\alpha(1-\alpha))$ and reduces to $-8t/9$ only at $\alpha = \frac{1}{2}$. □

Estimated constants and the confidence budget. Steps 1–7 treat L_w, L_r, L_a as fixed by conditioning on $\mathcal{C}$. When they are estimated, allocate the budget as $\delta = \delta_0 + \delta_1 + \delta_2 + \delta_3$ and replace (C) by the intersection of four events: $\mathcal{E}_{\delta_0}$ above, and one Hoeffding event per estimated constant. L_w and L_r are sample means of $\ell(\cdot, a_{\text{direct}})$ over the direct-wrong and direct-correct subsets, of sizes n_w and n_r, so $|\widehat{L}_\bullet - L_\bullet| \le L_{\max}\sqrt{\log(2/\delta_\bullet)/(2n_\bullet)}$ on the corresponding event. L_a is the fallback's expected loss, and is not estimated in the instantiation Table 8 reports, where the fallback is abstain and $\ell(\cdot, a_{\text{abstain}}) \equiv 1$; under the defer fallback of Appendix C.1 it is, and takes the same form. On the intersection, $\alpha_{\min} = (L_a - L_r)/(L_w - L_r)$ is perturbed by at most a factor $1 + O(\sum_\bullet |\widehat{L}_\bullet - L_\bullet|/d)$, and (D) is applied with β replaced by β minus that perturbation. Since n_w, n_r are of order n while the binding term in (D) is of order $\sqrt{t/m}$ with $m = \lfloor nq \rfloor$, the constant-estimation terms are dominated whenever (G) binds, which is why the main text states the threshold in the fixed-constant form.

A.4 Proposition 1 (information-theoretic lower bound)

The sign of the loss difference $\Delta(q)$ is determined by whether the bottom-q precision $\mu_w(q)$ exceeds $\alpha_{\min}$. Any test that determines this sign from n i.i.d. samples must in particular solve the binary hypothesis test $H_0 : \mu_w = \alpha_{\min}$ versus $H_1 : \mu_w = \alpha_{\min} + \beta$ using the $m = \lfloor nq \rfloor$ Bernoulli observations (wrong/right labels) in the rejected set.

Let $P_0 = \text{Bern}(p)$, $P_1 = \text{Bern}(p + \beta)$, with $p = \alpha_{\min}$. For any test φ, the Bretagnolle–Huber form of Le Cam's inequality gives

$$P_0^{\otimes m}(\varphi = 1) + P_1^{\otimes m}(\varphi = 0) \;\ge\; \tfrac{1}{2}\exp\big(-m\,\mathrm{KL}(P_0 \| P_1)\big).$$

If both type-I and type-II errors are at most δ, the left-hand side is at most 2δ, hence

$$m\,\mathrm{KL}(P_0 \| P_1) \;\ge\; \log\frac{1}{4\delta}.$$

A standard local Bernoulli KL bound gives, for $0 < \beta \le \frac{1}{2}\min\{p, 1-p\}$,

$$\mathrm{KL}\big(\text{Bern}(p) \,\|\, \text{Bern}(p+\beta)\big) \;\le\; \frac{2\beta^2}{p(1-p)}.$$

Combining the two displays gives

$$m \;\ge\; \frac{p(1-p)}{2\beta^2}\log\frac{1}{4\delta}.$$

It remains to relate $p(1-p)$ to $\alpha(1-\alpha)$. Under the proposition's assumption $\beta \le \frac{1}{2}\min\{p, 1-p\}$, we have $\alpha = p + \beta \le 3p/2$ and $1-\alpha = 1-p-\beta \ge (1-p)/2$; together with $\alpha \ge p$ and $1-\alpha \le 1-p$, this gives $\alpha(1-\alpha) \in [p(1-p)/2,\ (3/2)p(1-p)]$. Hence $p(1-p)$ and $\alpha(1-\alpha)$ differ by at most a factor of three, and substituting $m = \lfloor nq \rfloor$ yields (2) after adjusting the universal constant c. □

A.5 Theorem 3 (distribution-dependent partition lower bound)

The explicit per-cell requirement referred to in the theorem statement is as follows. Moreover, let σ^2 be an upper bound on the per-action conditional loss variance on any cell. Under Bernstein concentration on each cell, the empirical Π_1 estimator with κ-fold CV on n samples achieves cell-loss no worse than the global baseline $a^\star$ on cluster g (i.e. secures the corresponding $p_g\gamma_g$ contribution to (3)) with probability at least $1-\delta$ whenever $n \ge \frac{\kappa}{\kappa-1}\max\{8\log(2/\delta)/p_g,\ (16\sigma^2 + (8/3)L_{\max}\gamma_g)\log(4|\mathcal{A}|/\delta)/(p_g\gamma_g^2)\}$, where the first term is a Chernoff occupancy floor and the second the Bernstein concentration term.

For the population lower bound, condition on the cell G_g and observe that on each cell the partition router plays $a_g^\star$, the cell-optimal action, so its conditional loss is $\mathbb{E}[\ell(x, a_g^\star) \mid x \in G_g]$. Averaging over cells,

$$\Delta_{\text{partition}}(\mathcal{G}) = \sum_g p_g\big(\mathbb{E}[\ell(x,a^\star) \mid x \in G_g] - \mathbb{E}[\ell(x,a_g^\star) \mid x \in G_g]\big) = \sum_g p_g\gamma_g.$$

Each summand is non-negative because $a_g^\star$ minimizes conditional loss on G_g, so $\Delta_{\text{partition}}(\mathcal{G}) \ge \max_g p_g\gamma_g$, which is (3).

For the sample-complexity statement, condition on the partition and let $M_g \sim \text{Binom}(n_{\text{tr}}, p_g)$ be the realized training-fold count on cell G_g, with $n_{\text{tr}} = (\kappa-1)n/\kappa$ and mean $\bar{m}_g = n_{\text{tr}}p_g$. Because Bernstein's bound is convex in $1/m_g$, substituting $\bar{m}_g$ directly is not rigorous; we first lower-bound M_g by a multiplicative Chernoff inequality, $\mathbb{P}(M_g \ge \bar{m}_g/2) \ge 1 - \exp(-\bar{m}_g/8)$, then condition on this event with $\delta/2$ budget. On it, Bernstein's inequality applied to each action's empirical mean on cell G_g with variance at most σ^2 and range at most $L_{\max}$ gives

$$\mathbb{P}\big(|\widehat{L}_g(a) - L_g(a)| \ge t\big) \le 2\exp\left(-\frac{(\bar{m}_g/2)\,t^2}{2\sigma^2 + (2/3)L_{\max}t}\right).$$

Union-bounding over the $|\mathcal{A}|$ actions with the remaining $\delta/2$ budget, the empirical argmin's Bernstein deviation is at most $\gamma_g/2$ on every action whenever $\bar{m}_g \ge (16\sigma^2 + (8/3)L_{\max}\gamma_g)\log(4|\mathcal{A}|/\delta)/\gamma_g^2$, in which case $\widehat{L}_g(\widehat{a}_g^\star) \le \widehat{L}_g(a_g^\star) \le L_g(a_g^\star) + \gamma_g/2$ and therefore $L_g(\widehat{a}_g^\star) \le L_g(a_g^\star) + \gamma_g = L_g(a^\star)$, so the empirical Π_1 estimator's cell loss is no worse than the global baseline on cell G_g. Substituting $\bar{m}_g = (\kappa-1)np_g/\kappa$ and rearranging gives the stated bound. Strictly, the requirement is $\bar{m}_g \ge \max\{8\log(2/\delta),\ (16\sigma^2 + (8/3)L_{\max}\gamma_g)\log(4|\mathcal{A}|/\delta)/\gamma_g^2\}$, where the first term is the Chernoff floor on cell occupancy and the second the Bernstein concentration term, matching the explicit max in the theorem statement; the Bernstein term dominates whenever γ_g is bounded away from $L_{\max}$. □

A.6 Theorem 5 (held-out selection bound)

Notation. Let $M := |\mathcal{M}|$ be the number of candidate classes, $n_{\text{in}} := n_{\text{out}}/\kappa$ the size of one held-out inner fold, and $\ell \in [0, L_{\max}]$. For $m \in \mathcal{M}$ let $\widehat{R}_{\text{CV}}(m)$ be the inner κ-fold CV estimate and $\bar{R}_\kappa(m) := \mathbb{E}_{S'}[R(\widehat{\pi}_m(S'))]$ the expected risk of the estimator trained on a single inner-training fold of size $n_{\text{out}}(\kappa-1)/\kappa$. Write $m^\star := \arg\min_m \bar{R}_\kappa(m)$ and $\widehat{m} := \arg\min_m \widehat{R}_{\text{CV}}(m)$.

Step 1 (per-class deviation). Condition on the inner-training split, so that each fold-trained estimator is fixed and its held-out fold consists of n_{in} fresh samples with loss in $[0, L_{\max}]$. Hoeffding's inequality gives, for each m and each $t > 0$,

$$\mathbb{P}\big[\,|\widehat{R}_{\text{CV}}(m) - \bar{R}_\kappa(m)| > t\,\big] \;\le\; 2\exp\Big(-\tfrac{2n_{\text{in}}t^2}{L_{\max}^2}\Big). \tag{A}$$

Averaging over the κ folds preserves this order, so we use (A) for the CV average itself.

Step 2 (uniform event). Define, for $t > 0$,

$$\mathcal{U}(t) := \Big\{ \max_{m \in \mathcal{M}} \big| \widehat{R}_{\mathrm{CV}}(m) - \bar{R}_\kappa(m) \big| \le t \Big\}.$$

A union bound over the M classes applied to (A) gives

$$\mathbb{P}\big[\mathcal{U}(t)^c\big] \;\le\; 2M \exp\Big(-\tfrac{2n_{\mathrm{in}}t^2}{L_{\max}^2}\Big). \tag{B}$$

Step 3 (oracle inequality on $\mathcal{U}(t)$). On $\mathcal{U}(t)$, chaining the three inequalities $\bar{R}_\kappa(\widehat{m}) \le \widehat{R}_{\mathrm{CV}}(\widehat{m}) + t$, $\widehat{R}_{\mathrm{CV}}(\widehat{m}) \le \widehat{R}_{\mathrm{CV}}(m^\star)$ (by definition of $\widehat{m}$), and $\widehat{R}_{\mathrm{CV}}(m^\star) \le \bar{R}_\kappa(m^\star) + t$ yields

$$Z \;:=\; \bar{R}_\kappa(\widehat{m}) - \min_{m \in \mathcal{M}} \bar{R}_\kappa(m) \;\le\; 2t \qquad \text{on } \mathcal{U}(t). \tag{C}$$

Equivalently $\{Z > 2t\} \subseteq \mathcal{U}(t)^c$, so by (B), $\mathbb{P}[Z > 2t] \le 2M \exp(-2n_{\mathrm{in}}t^2/L_{\max}^2)$.

Step 4 (tail integration). Set

$$t^\star := L_{\max}\sqrt{\tfrac{\log(2M)}{2n_{\mathrm{in}}}},$$

the value at which the right-hand side of (B) equals 1. Since $Z \ge 0$, $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}[Z > z]\,\mathrm{d}z$; substituting $z = 2t$ gives $\mathbb{E}[Z] = 2\int_0^\infty \mathbb{P}[Z > 2t]\,\mathrm{d}t$. Split the integral at $t^\star$ and bound each piece by a different one of the two available bounds on $\mathbb{P}[Z > 2t]$:

$$\underbrace{2\int_0^{t^\star} \mathbb{P}[Z > 2t]\,\mathrm{d}t \;\le\; 2t^\star}_{\text{integrand} \le 1}, \qquad \underbrace{2\int_{t^\star}^\infty 2Me^{-2n_{\mathrm{in}}t^2/L_{\max}^2}\,\mathrm{d}t \;\le\; \frac{L_{\max}^2}{2n_{\mathrm{in}}t^\star} \;\le\; 2t^\star}_{\text{Gaussian tail, using } 2Me^{-2n_{\mathrm{in}}(t^\star)^2/L_{\max}^2} = 1},$$

where the middle bound uses $\int_{t^\star}^\infty e^{-ct^2}\mathrm{d}t \le e^{-c(t^\star)^2}/(2ct^\star)$ with $c = 2n_{\mathrm{in}}/L_{\max}^2$, and the last inequality is $L_{\max}^2/(4n_{\mathrm{in}}(t^\star)^2) \le 1$, i.e. $\log(2M) \ge \frac{1}{2}$, which holds for $M \ge 2$. Adding the two pieces,

$$\mathbb{E}[Z] \;\le\; 4t^\star \;=\; 2\sqrt{2}\,L_{\max}\sqrt{\tfrac{\log(2M)}{n_{\mathrm{in}}}} \;\le\; 4L_{\max}\sqrt{\tfrac{\log(2|\mathcal{M}|)}{n_{\mathrm{in}}}}, \tag{D}$$

which is (4).

Step 5 (refit transfer). Inequality (D) bounds the inner-CV estimand $\bar{R}_\kappa$. If refitting a selected class from inner-training size to the full outer-training size changes its expected risk by at most $\varepsilon_{\mathrm{stab}}(n_{\mathrm{out}}, \kappa)$, and the same holds for $m^\star$, then both sides of (D) move by at most that amount and the bound transfers to the refit estimator with an added $2\varepsilon_{\mathrm{stab}}(n_{\mathrm{out}}, \kappa)$. □

B Setup and implementation

B.1 Feature block specification

For HallusionBench and A-OKVQA the controller consumes a fixed scalar feature block of 41 dims (HallusionBench) or 39 dims (A-OKVQA), built entirely from the upstream Qwen2.5-VL prediction-time outputs (per-action margins, agreements, sampled-completion statistics, modality-bypass probes, and self-verification probes); the InternVL2.5-8B judge is used *only* to populate the semantic-risk channel of the loss matrix (Appendix B.2) and never enters the feature block. FOLIO uses a 12-dimensional simplified block constructed from the same Qwen2.5-VL outputs in text-only mode (described at the end of this subsection). SMS-Spam uses a 12-dimensional TF-IDF–based block (Appendix C.1). The TextVQA-OCR add-on uses a separate non-OCR feature block (exposed to Π_0, Π_1, Π_2; OCR-derived features go only to Π_3): three scalars (direct-confidence, question length in tokens, number of image classes) plus one-hot indicators (the first question token, the first 12 question tokens, tokens matching a small fixed text-key vocabulary, and the first 8 image-class names), assembled via a sparse `DictVectorizer` so the realized width is vocabulary-dependent. The Π_3 prior block additionally exposes OCR-derived signals (OCR confidence/margin, prediction length and digit/year flags, question/prediction overlap, training-frequency log) under the deterministic OCR-confidence gate. The hallucination/VQA block is organized into six groups, each of which expands to multiple scalars:

1. *Per-action confidence and margin* (7 dims). For each of the three non-abstain actions (direct, retrieve, defer): the greedy softmax margin and the mean per-token log-probability of the generated answer. Additionally, a context-aware margin for the retrieval action. These features capture first-order output-distribution signals for every candidate action.
2. *Cross-action agreement* (3 dims). Pairwise token-level agreement between the answers produced by direct, retrieve, and defer (`agree_dr`, `agree_de`, `agree_re`; the suffix letters are implementation identifiers and predate this paper's terminology).
3. *Question metadata* (2 dims). Question length in words and cosine similarity between question and context embeddings.
4. *Semantic similarity and benchmark categories* (8 dims on HallusionBench, 10 on A-OKVQA). Pairwise cosine similarities among the sentence embeddings of the three action-specific answers and between each answer and the question (6 pairs), plus benchmark-specific categorical one-hot indicators (HallusionBench: 2 dims, VD/VS category; A-OKVQA: 4 dims, question-type).
5. *Stochastic-sampling uncertainty* (8 dims, from $K{=}10$ sampled completions). Self-consistency count ratio (max-answer frequency divided by K), sample disagreement, Shannon semantic entropy over parsed distinct answers, predictive entropy from the first-token softmax, and means and standard deviations of the per-sample softmax margin and sequence log-probability across the K draws.
6. *Verification and modality-bypass probes* (13 dims on HallusionBench, 9 on A-OKVQA). Five scalars from a modality-bypass (image-blanked) probe applied to the vision benchmarks: a binary disagreement flag, softmax margin, sequence log-probability, $P(\text{yes}){-}P(\text{no})$, and top-1 token probability. Plus four scalars from a 3B self-verification re-ask ($P(\text{yes})$, $P(\text{no})$, their difference, and a binary text-contains-yes flag); on HallusionBench this is repeated under the 7B re-asker, contributing four further scalars ($P(\text{yes})$, $P(\text{no})$, their difference, and the binary text-contains-yes flag).

All features are computed once per sample on the upstream pipeline and represented as a dense matrix paired with the loss matrix. Features are standardized per outer-training fold via `sklearn.preprocessing.StandardScaler`, and all AUCs reported in Table 8 are computed with `sklearn.metrics.roc_auc_score` on the concatenation of outer-test folds so that the reported α_{emp} reflects the same data budget the controllers see.

The FOLIO 12-d block is built from groups 1–2 above (no $K{=}10$ stochastic, no blind-probe, no self-verification, no InternVL features, no question metadata) plus two predicted-label indicators (`pred_true`, `pred_uncertain`): per-action confidence margin and sequence log-prob (6 dims), three pairwise agreements (3 dims), the recover-context margin (1 dim), and the two indicator features (2 dims). This constrained block is what places FOLIO in the variance-bounded regime: Theorem 2's threshold $n_{\min} \propto \alpha(1{-}\alpha)/\beta^2$ depends on the AUC the feature block can extract, so a thinner block at small n instantiates the Thm 4 (ii) sub-case where Π_2's sign is uncertified.

B.2 Loss matrix construction

Loss components and per-benchmark labeling. The per-sample loss $\ell(x_i, a)$ is constructed from three operational quantities. Correctness $c(x_i, a) \in \{0, 1\}$ is evaluated against the benchmark gold label. Operational cost $k(x_i, a)$ is measured in action-cost units (1.0 for direct, 2.0 for retrieve, 2.3 for defer, 0.0 for abstain). Semantic risk $h(x_i, a) \in \{0, 0.5, 1\}$ is a per-action grounded-support label: $h{=}0$ (grounded), 0.5 (unverifiable), 1.0 (contradicted), with $h{=}0$ for a_{abstain} by construction. Labeling is benchmark-specific and run offline. *HallusionBench* and *A-OKVQA*: each (sample, action) candidate answer's most specific visual proposition is extracted and classified against image evidence by an InternVL2.5-8B judge as grounded / unverifiable / contradicted. *FOLIO* uses a rule-based proxy ($h{=}0.5$ if the action's answer disagrees with the gold label, $h{=}0$ if it agrees), keeping the loss matrix simple while still exposing wrong answers to a calibrated penalty; using a learned NLI judge here would conflate the controller's behaviour with the judge's quality on logical inference. *SMS-Spam* uses canonical-cost weighting without a separate judge (Appendix C.1).

Headline formula and weight robustness. The headline core loss is

$$\ell(x_i, a) = 1.0 \cdot (1 - c(x_i, a)) + 1.0 \cdot h(x_i, a) + 0.05 \cdot k(x_i, a),$$

which is the exact formula used for the four core benchmarks throughout this paper. Abstention therefore has loss 1.0, because it is treated as an always-incorrect zero-risk zero-cost action. Since this weight vector is a cost-sensitive operating point, Table 6 reruns the deployable $\Pi_0/\Pi_1/\Pi_2$ pool under local one-factor perturbations of these weights: $w_c \in \{0.75, 1.0, 1.25\}$, $w_h \in \{0.75, 1.0, 1.25\}$, and $w_k \in \{0.025, 0.05, 0.10\}$, varying one coordinate at a time around $(1, 1, 0.05)$. The TextVQA-OCR add-on instead uses answer loss $1-$soft-VQA accuracy because the experiment isolates OCR copying rather than semantic-risk/cost tradeoffs.

Diagnostic constants and numerical examples. The quantities $L_r := \mathbb{E}[\ell(x, a_{\text{direct}}) \mid c_{\text{direct}}{=}1]$, $L_w := \mathbb{E}[\ell(x, a_{\text{direct}}) \mid c_{\text{direct}}{=}0]$, and $L_a := \mathbb{E}[\ell(x, a_{\text{abstain}})]$ are computed as simple sample means over the respective subsets of the full core benchmark for the Π_2 viability diagnostic; they are global (not fold-specific) constants for each benchmark, so $\alpha_{\min} = (L_a - L_r)/(L_w - L_r)$ is a single number per benchmark that does not depend on the CV fold. For scale, the `always_direct` losses decompose under this formula as HallusionBench $1.0848 = (1 - 0.6685) + 0.7033 + 0.0500$, A-OKVQA $0.4138 = (1 - 0.8306) + 0.1944 + 0.0500$, and FOLIO $0.7520 = (1 - 0.5320) + 0.2340 + 0.0500$. Correctness c is gold-evaluated and therefore identical across judge choices; on HallusionBench and A-OKVQA the per-action risk h values entering Table 7 come from the InternVL2.5-8B judge from a different VLM family than the upstream Qwen2.5-VL policies, so the judge does not score outputs from its own family. On HallusionBench the population-mean per-action losses are $(1.085, 1.136, 1.104, 1.000)$ for (direct, retrieve, defer, abstain), so abstain becomes the best fixed action by a wide margin and the `fair_fixed_train` CV variant reliably recovers it; on A-OKVQA and FOLIO direct remains the best fixed action on outer-test. Table 7's Π_0 column reports the strict-CV-selected variant on each benchmark.

B.3 Family pool and baseline implementations

Our family pool is evaluated through a uniform `fit(X, L, C)`/`predict(X)` interface in the strict nested CV pipeline. Π_0 holds `always_direct` and a fair-fixed CV variant. Π_1 (KMeans) uses KMeans routers with $K \in \{4, 5, 6, 8\}$ on the standardized feature block, with per-cell action assigned by conditional loss argmin on outer-train and a global-best fallback for cells with fewer than three training samples. An alternative Π_1 (CART) realization uses `DecisionTreeClassifier` with `max_depth`$\in \{3, 4\}$ and `min_samples_leaf`$= 5$, fit on per-sample loss-argmin labels with the same per-leaf override and small-cell fallback (Table 5). Π_2 uses `HistGradientBoostingClassifier` (`max_depth`$\in \{3, 4\}$) and a calibrated logistic plug-in (C=0.3, referred to as `Selective-`C=0.3 elsewhere in this paper), with each family predicting a four-action argmin from the feature block.

We additionally evaluate two cost-sensitive adaptations of canonical learning-to-defer baselines through the same interface. The Mozannar–Sontag adaptation (Mozannar & Sontag, 2020) is a row-replicated multinomial logistic regression: each training row x_i is replicated $|\mathcal{A}| = 4$ times with the action index as the label and a sample weight $w_{i,a} = L_{\max,i} - L[i, a]$ encoding the cost margin, with L_2 regularization $C \in \{0.3, 1.0\}$; at test time we predict the softmax argmax. The Narasimhan adaptation (Narasimhan et al., 2022) is a post-hoc plug-in: one `HistGradientBoostingRegressor` per action (`max_iter`$= 200$, `max_depth`$\in \{3, 4\}$, `learning_rate`$= 0.05$) is trained on the per-action loss column, and the test policy is $\pi(x) = \arg\min_a \widehat{L}_a(x)$. Both are loss-matrix-compatible cost-sensitive instantiations rather than reproductions of the original human-deferral objectives; their detailed strict-CV results are in Appendix E.1.

B.4 Evaluation protocol and reproducibility

Strict nested CV. All headline tables use strict nested 5-fold-by-5-seed cross-validation with family selection on outer-train only, followed by a refit of the selected family on outer-train only and a single evaluation on outer-test. Non-strict variants that refit on outer-train-plus-outer-test are excluded from the canonical results: in pre-paper diagnostic runs on HallusionBench's best Π_2 family the non-strict refit produced losses inflated by $\sim$0.1 relative to the strict variant, the textbook overfit signature of refitting on the evaluation fold. Families, hyperparameter grids, and seeds are fixed; the strict outer CV is deterministic given the fixed seeds and pre-computed loss matrix.

Retrieve action. The a_{retrieve} action on HallusionBench, A-OKVQA, and FOLIO is generated by the same upstream model (Qwen2.5-VL-3B-Instruct) under a grounded-context prompting pass: a single forward call instructs the model first to enumerate scene context literally (objects, attributes, relations on visual benchmarks; key facts and entities on FOLIO) and then to commit to the question's answer conditional on that enumeration, with the final commitment extracted by deterministic regex from the response. There is no external retrieval corpus; the "retrieved" context is the model's own free-form description, so a_{retrieve} costs one extra forward pass over a_{direct} but draws on no information unavailable to the upstream model.

Hardware. All experiments run on AMD EPYC 7543 CPU cores and a single NVIDIA A100 GPU.

C PER-BENCHMARK DETAILS

C.1 SMS-SPAM: RESIDUAL-BOUNDED WITNESS FOR THEOREM 1

The SMS-Spam Collection (Almeida et al., 2011) is a 5,574-message English binary classification corpus (ham vs. spam). We use a deterministic 80/20 split (n_{train}=4460, n_{test}=1114; seed-0 NumPy permutation) and CPU-only TF-IDF features. The three deployable actions are: *direct* = TF-IDF (1,2)-gram (max_features 50,000, min_df 2) + calibrated LinearSVC (C=1.0, 3-fold sigmoid calibration), 0.9910 test accuracy; *retrieve* = TF-IDF (1,1)-gram (max_features 20,000, min_df 2) + KNN (k=10, cosine), 0.9811; *defer* = same TF-IDF (1,1)-gram + logistic regression (C=1.0), 0.9758. The 12-d scalar feature block is per-action top-1 minus top-2 probability, per-action top-1 log-probability, three pairwise prediction agreements, and 3 TruncatedSVD components fit on the SMS training split's (1,2)-gram TF-IDF matrix and applied to the held-out split. Loss uses canonical core weights $(1, 1, 0.05)$ and the same $\{1.0, 2.0, 2.3, 0.0\}$ action-cost vector as the other core benchmarks; abstain has zero operational cost k and (per the convention in Section 2.1) loss 1.0 under the canonical correctness weight.

This configuration is what makes SMS-Spam residual-bounded: the bigram + calibrated SVC direct is task-specialized, while the unigram logistic-regression defer is generic and not stronger on this task. We measure the four diagnostics from Section 2.2 and Theorem 1: $\mathbb{P}(R)$=0.009, oracle Π_2 gain =0.003, oracle Π_1 partition gain =0.000 (over a post-hoc $K \in \{2, 3, 4, 5, 6, 8\}$ KMeans sweep on an SVD-8 diagnostic embedding), and the local tail-margin condition of Lemma 1 fails: $\mathbb{P}(R) = 0.009$ caps the bottom-q precision by $\mu_w(q) \leq \mathbb{P}(R)/q \leq 0.03$ for $q \geq 0.3$, below $\alpha_{\min} = 0.95$ (abstain fallback) and $\alpha_{\min} = 0.089$ (defer fallback), so no rejector improves on a_{direct} asymptotically. We record this as $\beta \leq 0$ in Table 8 under the AUC scalar-proxy convention. All four are below the residual-bounded thresholds ($< 0.02, < 0.005, < 0.005$). Theorem 3 sharpens the Π_1 diagnostic: $\max_g p_g\gamma_g$=0 means every cluster's cell-best action coincides with the global-best a_{direct}, so $\pi_1^\star \in \Pi_0$ at the population level. Under strict nested 5-fold-by-5-seed CV, every Π_1 family (KMeans-$K \in \{4, 5, 6, 8\}$) collapses to direct ($\Delta = +0.0000$), the best Π_2 family is `SelectiveCalibrated_C0.3` at 0.0591±0.0000 ($\Delta = +0.0002$), and the two HGBC families lose by $\Delta \in \{+0.0007, +0.0011\}$ to the fixed action—the ordering the prediction expects, though every gap here is at the fourth decimal and none of them is resolved.

A simple structural lower bound on $\mathbb{P}(R)$ explains why this regime is rare under standard deferral setups. If defer's accuracy exceeds direct's by $\delta > 0$, then $\mathbb{P}(R) \geq \delta$ (the direct-wrong/defer-correct samples already form a δ-fraction). Under the canonical defer cost premium $0.05(c_{\text{defer}} - c_{\text{direct}}) = 0.065$, residual loss is bounded only when δ is comparable to or below this premium. The other three core benchmarks all have substantial residual mass: HallusionBench (defer-direct accuracy gap δ=0.037) and A-OKVQA (δ=0.034) violate the structural lower bound by defer's accuracy advantage; FOLIO ($\delta = -0.015$, defer slightly weaker) instead has $\mathbb{P}(R)$=0.175 driven by partition-level abstain dominance on its Uncertain subgroup. All three sit outside Theorem 1's scope; SMS-Spam, with a generic and weaker fallback that produces neither a defer-side accuracy advantage nor a partition-level abstain region, satisfies it. The witness is therefore not a property of the corpus alone but of the cost-tiered configuration in which the fallback is non-specialized—which is the deployment regime Theorem 1 is meant to characterize.

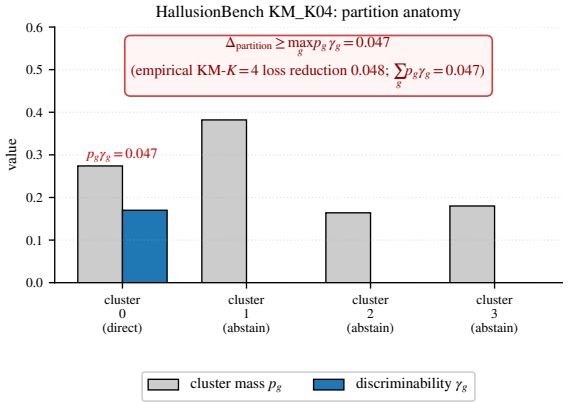

Figure 3: Cluster anatomy of HallusionBench KM-K=4, seed 42, witnessing Theorem 3. Under the canonical loss matrix the global best fixed action is abstain; only cluster 0 (p_0=0.274) carries positive discriminability by preferring direct (γ_0=0.170), contributing $p_0\gamma_0$=0.047. The empirical strict-CV KMeans-K=4 loss reduction is 0.048, within sampling noise of $\sum_g p_g\gamma_g$=0.047; the bound is empirically tight at K=4 on HallusionBench.

C.2 HallusionBench: instance-level wins on the high-signal Bernstein-viable regime

Under the InternVL2.5-8B judge, HallusionBench at the full n=920 corpus is comfortably above the Bernstein threshold ($n^{\text{Bern}}_{\min}$=23, with $n\beta^2$=202; Theorem 2); both Π_1 (best KMeans-K=8, 0.091 loss reduction) and Π_2 (best HGBC-`md3`, 0.103 loss reduction) beat the fixed action, with Π_2 edging Π_1 by ≈ 3 seed-sd of Π_2. The deployable winner is therefore Π_2; Π_1 remains a useful lower-complexity diagnostic for Theorem 3. Figure 3 reports the partition anatomy at the simplest router (K=4, seed 42): InternVL2.5-8B's stricter judging of direct-wrong samples makes a_{abstain} the global best fixed action under the canonical loss matrix (mean direct loss 1.085 vs. abstain's 1.000), so by the definition $\gamma_g = \mathbb{E}[\ell(x,a^\star) - \ell(x,a^\star_g) \mid x \in G_g]$ a cluster contributes positively only when its cell-best action differs from $a^\star$=a_{abstain}. One cluster carries essentially all the partition gain at K=4: cluster 0 (p_0=0.274) prefers a_{direct} with γ_0=0.170, contributing $p_0\gamma_0$=0.047; the remaining three clusters all prefer a_{abstain} and contribute zero, so $\max_g p_g\gamma_g = \sum_g p_g\gamma_g = 0.047$. The strict-CV empirical KM-K=4 loss reduction is 0.048 ± 0.005 relative to the `fair_fixed_train` baseline of 1.000, within sampling noise of $\sum_g p_g\gamma_g$: Theorem 3's lower bound is empirically tight at K=4. The instance-level recoverable margin (HGBC-`md3` at 0.103) exceeds this partition ceiling, so Π_2 wins on HallusionBench by Theorem 4 item (iii)'s ceiling comparison, with the partition diagnostic confirming that Π_1's gain is concentrated in a single direct-preferring subgroup rather than diffused across the partition.

C.3 A-OKVQA: Π_1/Π_2 boundary

A-OKVQA at full n=1145 is the deployable Π_2 winner (Table 7, Selective C=0.3 at -0.033 versus best Π_1 KMeans-K=4 at -0.024), but without empirical separation. The ceilings are not close—$C_{\Pi_2} = q^\star\beta(L_w - L_r) \approx 0.094$ against $C_{\Pi_1} = \max_K \sum_g p_g\gamma_g \approx 0.039$, so Theorem 4 (iii) predicts Π_2 cleanly—yet the recovered empirical margins are within one seed-sd of each other and inner-CV splits Π_1:Π_2 as 10:15 over 25 outer-folds (Appendix E.3).

C.4 A-OKVQA: gold-rationale Π_3 oracle (non-deployable)

Because these rationales were collected after the answers and ask workers to explain why the correct answer is correct (Schwenk et al., 2022), they are answer-derived. We therefore treat the following as a gold-rationale *oracle experiment*: it measures the upper-bound value of an answer-derived explanatory channel, which no deployment can obtain at prediction time.

A-OKVQA (n=1145), best Π_2 per block	loss	std	Δ vs Π_3
Canonical scalar block (39-d), no rationale access			
Π_0: always_direct	0.4138	0.0000	+0.0584
Π_1: KMeans-K=4	0.3902	0.0085	+0.0348 (4.0 seed-sd)
Π_2: SelectiveCalibrated C=0.3	0.3805	0.0032	+0.0251 (6.8 seed-sd)
Π_2 *given three learned rationale representations*			
Π_2 HGBC-md3 + exact-match (40-d)	0.3749	0.0052	+0.0195 (3.5 seed-sd)
Π_2 HGBC-md3 + text-overlap stats (45-d)	0.3737	0.0045	+0.0183 (3.7 seed-sd)
Π_2 HGBC-md3 + TF-IDF top-200 (239-d)	0.3847	0.0043	+0.0293 (6.2 seed-sd)
Π_3*-style gold-rationale oracle (not a deployable feature)*			
Π_3 oracle: rationale gate + KM-K=4 fallback	**0.3554**	**0.0019**	**0** (**oracle**)

Table 3: Gold-rationale oracle decomposition on A-OKVQA (5-seed strict CV). **Top:** scalar-only deployable families. **Middle:** Π_2 HGBC given three learned representations of the rationale channel; note that the widest of them (TF-IDF, 239-d) is worse than the plain scalar Π_2 at 0.3805, so more rationale information does not monotonically help. **Bottom:** the deterministic gold-rationale gate plus KMeans fallback. None of the three augmented Π_2 variants closes the gap to the oracle gate (3.5–6.2 seed-sd above). Each row reports the lowest-mean Π_2 family within its block, so the table is a best-of comparison on both sides.

We define a parameter-free binary gate $\rho(x)$ that fires when any content word of the model's direct answer appears in any rationale of x. On A-OKVQA ρ fires on $806/1145{=}70.4\%$ of samples with empirical direct-correct rate 94.4% inside the fired set (vs. base rate 83.1%); on the remaining 29.6% the rate collapses to 56.0%, giving AUC 0.784. The Π_3-style family plays a_{direct} when $\rho{=}1$ and uses a KMeans-$K{=}4$ fallback on the non-fire subset, with no tunable hyperparameters.

Under strict 5-fold-by-5-seed CV Table 3 reports the run: the oracle reaches 0.3554 ± 0.0019, beating the best scalar Π_1 at 0.3902 ± 0.0085 by -0.0348 and the best scalar Π_2 at 0.3805 ± 0.0032 by -0.0251, gaps of about 4 and 7 seed-sd units, respectively. The result is numerically consistent with the region where Π_3 wins in Appendix D.3; the gate's AUC matches the high-prior-signal region.

C.5 FOLIO: VARIANCE-BOUNDED WITNESS FOR THEOREM 4 (II)

FOLIO (Han et al., 2024) is a 203-sample first-order-logic NLI benchmark in which each example consists of 1–7 premises and a conclusion, with a 3-way label (True / False / Uncertain) at $\approx$ 35:30:34%. We use the official validation split in full ($n{=}203$) and the same Qwen2.5-VL-3B-Instruct (a_{direct}, a_{retrieve}) and Qwen2.5-VL-7B-Instruct (a_{defer}) backbones in text-only mode, with rule-based per-action risk ($h{=}0.5$ if action's answer is wrong, $h{=}0$ otherwise). The 12-d feature block is per-action confidence margin, sequence log-prob, recover-context margin, three pairwise agreements, and two indicator features for the model's predicted-label class.

The full-n quantities are $\alpha_{\text{emp}}{=}0.687$, $\alpha_{\min}{=}0.633$, so $\beta \approx 0.0528$ (displayed as 0.053 to three decimals throughout) and $n^{\text{Bern}}_{\min}{=}1898$ from the leading-order formula at the unrounded β. FOLIO is therefore $\approx 9.4\times$ below the Bernstein threshold: Theorem 2 does *not* certify the sign of the empirical Π_2 loss difference, and Theorem 4 item (ii) predicts that whenever $\max_g p_g\gamma_g > 0$ a partition router strictly improves on Π_0 while Π_2 may have wrong sign. Both predictions hold empirically: the deployable Π_1 winner KMeans-$K{=}6$ reaches $0.7195{\pm}0.007$ versus Π_0's 0.7520 (a 0.0325 reduction, 4.5 seed-sd of Π_1); the deployable Π_2 winner Selective-$C{=}0.3$ reaches 0.7546 ± 0.009, $+0.0027$ worse than Π_0. That difference sits well inside its own seed-sd, which is the point: at $n \ll n^{\text{Bern}}_{\min}$ Theorem 4 (ii) declines to certify the sign, and the data do not supply one either.

The partition's anatomy is unbalanced and concentrated. The KMeans clustering recovers eight clusters at $K{=}8$; only three carry positive discriminability (the three clusters whose cell-best action is not the global best fixed action), with $p_g\gamma_g$ contributions $0.006, 0.012, 0.047$ summing to $\sum_g p_g\gamma_g{=}0.065$. The empirical KM-$K{=}6$ loss reduction 0.0325 is below this oracle ceiling because (i) strict-CV refits KMeans on each outer-train fold of size ~ 162, smaller than the full-pool seed-42 partition, and (ii) the ceiling assumes oracle cell-best actions whereas the empirical estimator must learn them. Π_2's failure to separate from Π_0 here is structurally driven: with $\beta{=}0.053$ the canonical Π_2 asymptotic ceiling is only $q^\star\beta(L_w{-}L_r) \approx 0.024$, which by Theorem 2 requires $n \geq 1898$ to

be reliably recovered; at n=203 the variance overhead exceeds the ceiling. The testable half of the prediction—that Π_1 wins—is what the protocol returns; the Π_2 half is a statement that no sign should be certifiable, and none is.

D CONTROLLED SYNTHETIC VALIDATION

We complement the per-benchmark results with three controlled synthetics that hold all but one structural variable fixed and sweep only the relevant diagnostic: Section D.1 sweeps n across the Bernstein viability threshold of Theorem 2; Section D.2 varies instance-level signal at fixed partition gap to test the Π_1/Π_2 phase transition (Thm 4 (iii)); Section D.3 varies an external prior channel's informativeness to characterize when Π_3 wins.

D.1 BOTTOM-q PRECISION-ESTIMATOR CROSS-THRESHOLD FOR THEOREM 2

The four core benchmarks place a single point on each side of the Bernstein threshold (FOLIO below, HallusionBench/A-OKVQA above), but they cannot show how sign correctness behaves *as* n crosses $n_{\min}$. The synthetic in Section D.2 sweeps Π_1/Π_2 instead and stays above threshold throughout. We therefore include a dedicated cross-threshold experiment on the same bottom-q precision estimator that Theorem 2 analyses (not the full Π_2/CV pipeline; see Appendix D.2 for that).

At fixed $(\alpha, q, \delta) = (0.75, 0.3, 0.05)$, we sweep $n \in \{20, 30, 50, 80, 130, 200, 320, 500, 800, 1300, 2000, 3200, 5000, 8000\}$ and $\beta \in \{0.05, 0.10, 0.20\}$ (so $n_{\min} \in \{1844, 461, 115\}$). For each (n, β), we draw 4000 replications of the bottom-q precision estimator (each replication: $m = \lfloor nq \rfloor$ i.i.d. Bernoulli(α) draws under the local tail-margin convention) and record the empirical sign-correctness rate $\Pr[\,\mathrm{sign}(\widehat{\Delta}(q)) = \mathrm{sign}(\Delta(q))\,]$. Theorem 2 predicts this rate is at least $1 - \delta = 0.95$ once $n \geq n_{\min}(\beta)$.

Figure 4 shows the resulting rates. For each β, the empirical curve crosses the $1 - \delta$ target at a sample size at or somewhat below $n_{\min}(\beta)$, confirming the qualitative prediction (sign certified for $n \gtrsim n_{\min}$, uncertified below) and showing scaling consistent with order-tight β^{-2} separation. The displayed grid's first empirical crossings sit at $n/n_{\min} \in [0.43, 0.69]$, the constant-factor gap typical between Bernstein-derived sufficient conditions and the actual finite-sample saturation point; Proposition 1 rules out improvement in the leading order, not in constants. FOLIO's empirical wrong-sign Π_2 lands in the curve's variance-bounded regime: its $n/n_{\min}$=203/1898=0.107 ratio sits below all three empirical-saturation crossings, so the wrong-sign observation is robust to the constant slack of the Bernstein sufficient condition. The synthetic uses α=0.75 versus FOLIO's α_{emp}=0.687, so the mapping is qualitative; HallusionBench at n=920, β=0.469 sits well past saturation.

D.2 Π_1/Π_2 PHASE TRANSITION FOR THEOREM 4

The data-generating process has four latent clusters of mass $(0.35, 0.25, 0.20, 0.20)$ in $\mathbb{R}^6$ with a residual heterogeneity that gives Π_1 a fixed population partition gap of $\max_g p_g \gamma_g \approx 0.08$. A scalar knob bk $\in \{0, 0.5, 1.0, 1.6, 2.4, 3.5\}$ controls how much smooth continuous signal the features carry about c_{direct}: at bk $= 0$ the only information is cluster identity, and as bk grows a continuous feature is added that makes the discriminative signal increasingly extractable by a regularized LR classifier. We sweep $n \in \{150, 300, 600, 1200, 2400, 4800\}$ crossed with bk, run three seeds per cell with Π_1 realized as KMeans-K=4 and Π_2 realized as a regularized LR softmax (a deliberately compact subset of the main-paper Π_2 pool, isolating capacity-realization effects), and declare the winner class. Every cell in the sweep sits comfortably above the Bernstein viability threshold of Theorem 2 (since β stays in $[0.37, 0.62]$ throughout), so this experiment is *not* a Theorem 2 validation. It is a direct test of the population-versus-empirical clause in item (iii) of Theorem 4: when the asymptotic ceilings $C_{\Pi_2} > C_{\Pi_1}$ at every cell, does the empirical winner still track the comparison once a finite-capacity Π_2 family is plugged in? Figure 5 plots the resulting winner map.

The result is precisely the population-versus-empirical gap that the second clause of Theorem 4 (iii) anticipates. The asymptotic instance-level ceiling $C_{\Pi_2} = q^\star \beta (L_w - L_r)$ exceeds the partition gap 0.08 at every cell since $\beta \in [0.37, 0.62]$ is large throughout, so the population-optimal clause of Thm 4 (iii) predicts Π_2. The empirical winner does not always match: at bk $= 0$, when the only source of margin is discrete cluster identity, Π_1 beats Π_2 at every $n \geq 300$ because Π_1's per-cluster

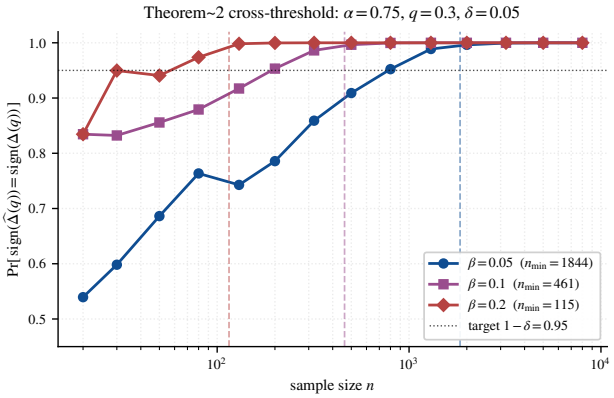

Figure 4: Bottom-q precision-estimator cross-threshold. At each β, the empirical sign-correctness rate (over 4000 replications per cell) reaches the target $1-\delta=0.95$ (dotted black) at sample sizes consistent with, and somewhat below, the predicted threshold $n_{\min}(\beta)$ (dashed vertical lines). The empirical curves are consistent with the order-tight β^{-2} scaling guaranteed by Proposition 1 but show the constant slack typical of Bernstein-style sufficient conditions; the synthetic does not run the full Π_2/CV pipeline and so does not directly verify constant-tightness in deployed controllers. Above each $n_{\min}$ the rate saturates at ≈ 1; below, it degrades smoothly as predicted by the variance-bounded sub-case of Theorem 4 (ii).

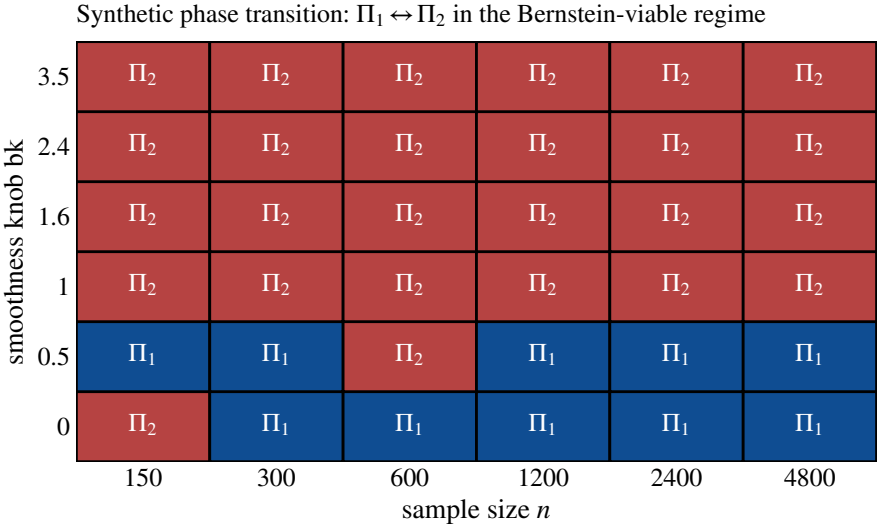

Figure 5: Controlled synthetic validation of Theorem 4 item (iii). Each tile is one (n, bk) cell colored by the empirical winner class (blue $=\Pi_1$, red $=\Pi_2$). The partition gap is held fixed at $\max_g p_g\gamma_g \approx 0.08$; the smoothness knob bk on the y-axis is the only manipulated variable, and the resulting empirical margin β stays in $[0.37, 0.62]$ across all cells, well above the Bernstein viability threshold of Theorem 2. At low bk the only signal is cluster identity and Π_1's direct per-cluster argmin beats Π_2's regularized LR softmax; at higher bk the added smooth feature gives Π_2 enough instance-level margin to overtake the partition gap. The empirical winner boundary thus realizes the direct comparison between instance-level margin and partition gain that Theorem 4 item (iii) predicts.

argmin recovers the cell-best action sharply while the regularized LR family under-realizes the same discrete signal. As bk grows the LR family's recoverable margin rises (closer to its asymptotic ceiling), and once it overtakes the fixed partition gain the empirical winner flips to Π_2 and stays there: Π_1 wins $10/12$ cells for bk $\in \{0, 0.5\}$ and Π_2 wins all $24/24$ cells for bk ≥ 1, with the two Π_2 wins at $(n{=}150, \text{bk}{=}0)$ and $(n{=}600, \text{bk}{=}0.5)$ within 0.003 of Π_1 (multi-seed noise). The synthetic

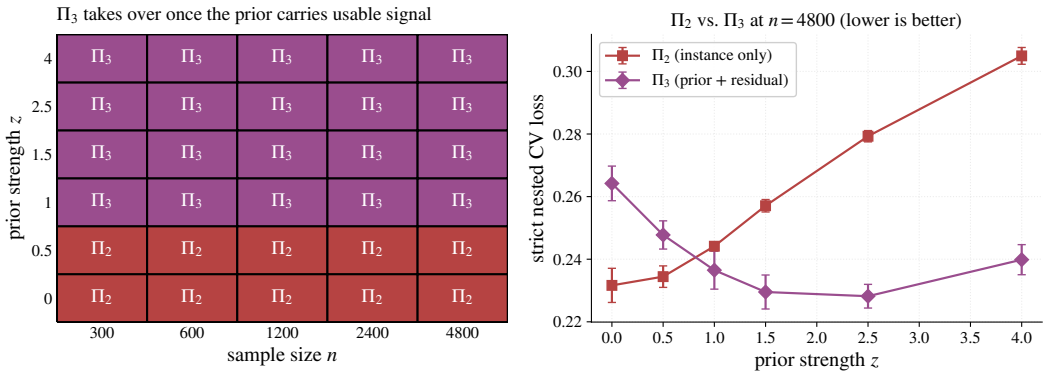

Figure 6: Controlled synthetic validation of Π_3, with the partition signal weakened ($|\mathrm{bump}| \leq 0.3$) and the Π_2 smooth-signal knob fixed at $\mathrm{bk}{=}1.6$ so Π_2 sits comfortably in its high-signal regime. **Left:** empirical winner tile plot across (n, z); each cell is a strict nested 5-fold CV outcome colored by the winning class (red$=\Pi_2$, purple$=\Pi_3$). At $z \in \{0, 0.5\}$ the prior is uncorrelated or only weakly correlated with correctness and Π_2 wins every cell (10/30); at $z \geq 1.0$ the prior carries enough signal about correctness that prior-gate decisions are more reliable than anything Π_2 can extract from X, and Π_3 wins every cell (20/30). **Right:** at $n{=}4800$, the Π_2 loss *grows* from 0.23 to 0.30 as z increases (the correctness logit becomes increasingly z-dominated and X correspondingly less informative), while the Π_3 loss falls from 0.26 at $z{=}0$ to the 0.23–0.24 band once $z \geq 1.0$. The separation between the two curves is the Π_3 gain.

therefore confirms that the empirical winner tracks the asymptotic-ceiling comparison only to the extent that the deployed Π_2 family realizes its ceiling. We did not sweep a higher-capacity Π_2 family such as HGBC here, so where the bk crossing falls for a richer family is untested.

D.3 Π_3 ORTHOGONAL-CHANNEL SWEEP

Π_3 can strictly improve on lower classes only when a label-free external prior channel carries information about correctness beyond what the Π_2 feature block provides. To characterize this condition directly, we sweep the prior's informativeness in a controlled DGP.

Inputs are drawn from the same four-cluster Gaussian as in Appendix D.2, with the partition weakened ($\mathrm{bump}{=}(+0.3, +0.3, +0.3, -0.3)$) and the Π_2 smooth-signal knob fixed at $\mathrm{bk}{=}1.6$, so Π_2 sits comfortably in its high-signal regime. We add a hidden prior scalar $z \sim \mathcal{N}(0, 1)$ drawn independently of X, with correctness logit $\mathrm{logit}(c_{\mathrm{direct}}{=}1 \mid X, z) = \mathrm{bk}\, s(X) + z_{\mathrm{strength}} \cdot z + \mathrm{bump}(g)$, where $s(X)$ is the unit-norm linear signal from Appendix D.2. The feature block exposed to Π_0, Π_1, Π_2 is X only; z is exposed exclusively to Π_3. Because z is independent of X, no amount of Π_2 training on X can recover the information z carries about correctness, regardless of n.

Π_3 is implemented as the Definition 1 prior-gated controller with an explicit high-confidence gate: $\Pi_3(x, z){=}a_{\mathrm{direct}}$ if $z > \tau$, $\Pi_3(x, z){=}a_{\mathrm{defer}}$ if $z < -\tau$, and $\Pi_3(x, z){=}r(x)$ otherwise, with $\tau{=}1.0$ so that the gate fires on the top/bottom $\approx 16\%$ of the prior and leaves the middle $\approx 68\%$ to the fallback. The asymmetric $\mathrm{bump} = (+0.3, +0.3, +0.3, -0.3)$ keeps the partition discriminable so that the synthetic does not become trivially Π_3-dominated by collapsing the Π_1 ceiling to zero. The fallback controller $r \in \Pi_2$ is a regularized multinomial logistic regression fit only on training samples with $|z_{\mathrm{tr}}| \leq \tau$. We cross $n \in \{300, 600, 1200, 2400, 4800\}$ with $z_{\mathrm{strength}} \in \{0, 0.5, 1.0, 1.5, 2.5, 4.0\}$, 3 seeds per cell, under the same strict 5-fold CV protocol.

The winner map (Figure 6 left) partitions the sweep into two regimes: when $z_{\mathrm{strength}} \leq 0.5$ the prior is too weak to justify the gating cost and Π_2 wins uniformly in n; when $z_{\mathrm{strength}} \geq 1.0$ the prior is accurate enough that Π_3's gated decisions dominate and Π_3 wins uniformly in n. The transition is crisp: there is no "Π_3 wins at large n but not small n" axis, because Π_3's advantage comes from additional information rather than from more samples. The right panel anchors the mechanism on Π_3: at $n{=}4800$, Π_3's loss falls from 0.264 at $z_{\mathrm{strength}}{=}0$ to the 0.23–0.24 band once $z_{\mathrm{strength}} \geq 1.0$, showing that the gate becomes useful only after the prior carries enough signal;

Π_2's loss rises from 0.23 to 0.30 by construction of the DGP, since fixing $\mathrm{bk}{=}1.6$ while raising z_{strength} makes X a relatively weaker driver of correctness, so the rising curve is a control rather than a finding. The displayed curves cross between $z_{\mathrm{strength}} = 0.5$ and 1.0 (linear interpolation at $n{=}4800$ gives $z_{\mathrm{strength}} \approx 0.82$), the operating point at which the prior carries enough signal that the certainty-on-the-tails gate outweighs the cost of routing the un-gated middle band to the fallback. The orthogonality $z \perp X$ is the cleanest possible setting for Π_3 and overstates the advantage available to a real channel that shares partial information with X; the TextVQA-OCR experiment in Section 3 is the deployable partial-overlap test, and the A-OKVQA oracle of Appendix C.4 agrees numerically.

E Baselines, ablations, and CV diagnostics

E.1 Cost-sensitive learning-to-defer adaptations

As a robustness check, we instantiate families from outside our own pool and ask whether they show the same qualitative failures and successes. We run two cost-sensitive adaptations of canonical learning-to-defer baselines under the identical strict nested CV protocol: a softmax surrogate adapted from Mozannar and Sontag (Mozannar & Sontag, 2020), and a post-hoc plug-in estimator adapted from Narasimhan et al. (Narasimhan et al., 2022). Both sit in Π_2 (instance-level learned controllers) by construction. The Mozannar variant is implemented as a row-replicated multinomial logistic regression with per-row weights $w_{i,a} = L_{\max,i} - L[i,a]$; the Narasimhan variant trains one HistGradientBoosting regressor per action $\widehat{L}_a(x)$ and selects $\pi(x) = \arg\min_a \widehat{L}_a(x)$ at test time. These are not drop-in reproductions of the original human-deferral objectives; they are loss-matrix-compatible implementations designed to test whether other Π_2-style learners encounter the same regimes. We sweep a small hyperparameter grid on each ($C \in \{0.3, 1.0\}$ for Mozannar, `max_depth`$\in \{3, 4\}$ for Narasimhan) and run them through the same 5-fold-by-5-seed strict CV pipeline as our own families. Table 4 reports the results.

Two limits bound what the comparison shows. On HallusionBench ($n{=}920$, Bernstein-viable), Narasimhan-`md3` (0.8990 ± 0.007) ties our HGBC-`md3` (0.8970 ± 0.004) within 0.3 joint seed-sd, while Mozannar-$C{=}0.3$ (0.9203 ± 0.005) trails by ≈ 3.6 joint seed-sd; the spread reflects implementation-level differences (post-hoc plug-in vs. row-replication) within a single Π_2 class, not a class-selection signal. Our Π_1 KMeans-$K{=}8$ (0.9093 ± 0.003) beats Mozannar by ≈ 1.9 joint seed-sd but trails Narasimhan within ≈ 1.4 joint seed-sd, consistent with HallusionBench sitting just inside the high-signal Π_2-favorable region. On A-OKVQA (high-signal boundary), Narasimhan-`md3` reaches 0.3756 ± 0.002, numerically below our Π_2 Selective at 0.3805 ± 0.003 by ≈ 1.4 joint seed-sd. Narasimhan lands 0.005 below our canonical-pool winner Selective. Both are Π_2-class, so the ordering between them says nothing about class selection, and it does say that our Π_2 family is not the best available one at this ceiling; Mozannar at 0.3877 shows that not every Π_2 implementation realizes the ceiling equally, an intra-Π_2 spread that does not flip the class-selection conclusion. On FOLIO (variance-bounded, $n{=}203 \ll n_{\min}{=}1898$), Mozannar-$C{=}1.0$ (0.7478 ± 0.008) and Narasimhan-`md3` (0.7705 ± 0.021) trail our KMeans-$K{=}6$ partition at 0.7195 ± 0.007 by ≈ 2.7 and ≈ 2.3 joint seed-sd respectively, matching the Π_1-wins-Π_2 sub-case of Theorem 4 (ii); both L2D adaptations are Π_2-class by construction and inherit the $n < n_{\min}^{\mathrm{Bern}}$ uncertified-sign behaviour the theory predicts. The pattern across the three benchmarks tracks the theory: when $n \geq n_{\min}^{\mathrm{Bern}}$, multiple independent Π_2 implementations cluster within seed noise; when $n \ll n_{\min}$, partition routers strictly dominate the Π_2 realizations we tested, including the two L2D adaptations outside our main family pool. The adapted L2D baselines are not a proof that the original L2D objectives would behave identically under every implementation choice; they show that the observed regime classification is robust beyond our HGBC family.

E.2 Ablations: CART partition router, loss-weight sensitivity

Two ablations probe the robustness of the regime map. The first asks whether the Π_1 winners depend on KMeans. We add a second, structurally different Π_1 realization: a shallow CART partition router (`DecisionTreeClassifier` at `max_depth`$\in \{3, 4\}$) whose splits are fit on the per-sample loss-argmin label and whose leaves are re-assigned the cell-wise loss-argmin action at fit time (Appendix B.3). Table 5 reports the per-benchmark numbers. On HallusionBench, KMeans-$K{=}8$ (0.9093) and CART-d4 (0.9104) tie within seed noise, both losing to the per-class Π_2 winner, so

Benchmark	Mozannar (best)	Narasimhan (best)	Best (ours)	Verdict
HallusionBench	0.9203 ± 0.005	0.8990 ± 0.007	$\mathbf{0.8970 \pm 0.004}$ (Π_2)	Narasimhan ties HGBC (0.3 sd); Mozannar trails (3.6 sd)
A-OKVQA	0.3877 ± 0.004	0.3756 ± 0.002	$\mathbf{0.3805 \pm 0.003}$ (Π_2)	Narasimhan lower by 1.4 sd; intra-Π_2
FOLIO	0.7478 ± 0.008	0.7705 ± 0.021	$\mathbf{0.7195 \pm 0.007}$ (Π_1)	ours wins (2.7/2.3 joint sd)

Table 4: Cost-sensitive adaptations of two learning-to-defer baselines (Mozannar–Sontag 2020, Narasimhan et al. 2022) evaluated under the identical strict nested 5-fold-by-5-seed CV protocol used in this paper. Numbers are mean loss $\pm$ seed std; lower is better. Bold marks our deployable per-benchmark winner from Table 7. Both adapted L2D baselines sit firmly in Π_2 by construction, so the comparison is intra-Π_2 on HallusionBench/A-OKVQA and inter-class on FOLIO. They recover the Π_2-class result on HallusionBench (Narasimhan ties our HGBC within 0.3 joint seed-sd; Mozannar trails by 3.6 joint seed-sd) and on A-OKVQA (Narasimhan numerically below our Selective by 1.4 joint seed-sd, intra-Π_2 variation), but trail our Π_1 winner on FOLIO by 2.7 and 2.3 joint seed-sd respectively, the variance-bounded sub-case (Thm 4 (ii)) where instance-level estimation is unreliable and partition routing is the regime-correct class. SMS-Spam is excluded because in its residual-bounded regime (Theorem 1) L2D wrappers reduce to direct by construction.

Benchmark	Π_0 best	KMeans Π_1 best	CART Π_1 best	verdict
HallusionBench	1.0000	0.9093 ± 0.003	0.9104 ± 0.004	both beat Π_0; tied within noise
A-OKVQA	0.4138	0.3902 ± 0.009	$\mathbf{0.3742 \pm 0.005}$	both beat Π_0; CART better
FOLIO	0.7520	$\mathbf{0.7195 \pm 0.007}$	0.7981 ± 0.042	KMeans wins; CART overfits

Table 5: Non-KMeans Π_1 realization: shallow CART partition routers under the identical strict nested 5-fold-by-5-seed CV protocol. Bold marks the Π_1 winner among KMeans/CART. On HallusionBench KMeans and CART tie within seed noise. On A-OKVQA CART-d4 at 0.3742 falls ≈ 1.1 joint seed-sd below the canonical Π_2 winner Selective at 0.3805, so the choice of partition family, not only the class, can decide this benchmark. On FOLIO, the depth-bounded tree overfits at n=203, consistent with Theorem 3's required per-leaf sample budget not being met; KMeans's fixed K acts as an implicit regularizer.

HallusionBench remains a clean Π_2 win regardless of partition method. On A-OKVQA the picture is more interesting: CART-d4 reaches 0.3742 ± 0.005, numerically below the canonical Π_2 Selective at 0.3805 ± 0.003 by ≈ 1.1 joint seed-sd. This is within seed noise. We retain Selective as the canonical Π_2 winner because the pool was fixed in advance, but the honest reading of A-OKVQA after this ablation is a Π_1/Π_2 tie rather than a clean Π_2 win. It is exactly the second clause of Theorem 4 (iii) at work: when C_{Π_1} and C_{Π_2} are commensurate, finite-capacity family realization—here the choice between an axis-aligned CART and a centroid-based KMeans—can flip the empirical winner between supportable adaptive classes. On FOLIO the depth-bounded CART overfits at n=203 and lands at 0.7981 ± 0.042, $+0.046$ *worse* than Π_0; KMeans-K=6 at 0.7195 remains the Π_1 winner. CART's FOLIO failure is the partition-router-side analogue of Thm 4 (ii)'s Π_2 failure: at n=203, depth-3-or-4 CART leaves carry too few samples for the empirical per-leaf argmin to track its cell-best action, consistent with the per-leaf sample budget required by Theorem 3 not being met; KMeans's fixed K is an implicit regularizer that the depth-bounded tree, with its more flexible leaf structure, lacks.

The second ablation addresses the cost-sensitivity of the combined loss. Classical reject-option and cost-sensitive evaluation explicitly depend on operating costs (Chow, 2003; Drummond & Holte, 2006), so a class-selection result should not hinge on a single arbitrary weight vector. We therefore rerun the deployable $\Pi_0/\Pi_1/\Pi_2$ family pool under seven local perturbations of the canonical core weights: $(1, 1, 0.05)$, correctness weight 0.75 and 1.25, risk weight 0.75 and 1.25, and cost weight 0.025 and 0.10, always with the same strict 5-fold-by-5-seed protocol. Table 6 summarizes the winner counts. Two of the three benchmarks are fully stable: A-OKVQA Π_2 $(7/7)$ and FOLIO Π_1 $(7/7)$. HallusionBench is the boundary case: the canonical-weights winner is Π_2 $(5/7)$, but two perturbations (`correct_high`, `risk_low`) swap the winner to Π_1 by less than 0.005, reflecting the fact that the canonical-weights gap between Π_1 (KM-K=8 at 0.909) and Π_2 (HGBC-`md3` at 0.897) is only ≈ 3 seed-sd of Π_2. The qualitative regime map is therefore robust on A-OKVQA and FOLIO, with HallusionBench's Π_1/Π_2 boundary slightly weight-dependent: the operating-cost neighborhood swept here, which spans $\pm 25\%$ perturbations of the correctness and risk weights and a $2\times$ range of cost weights around the canonical setting, leaves the deployable winner unchanged on A-

Benchmark	canonical class	winner counts over 7 weights	min winner margin	interpretation
HallusionBench	Π_2	Π_2:5, Π_1:2	0.0047	Π_1/Π_2 boundary
A-OKVQA	Π_2	Π_2:7	0.0049	stable Π_2
FOLIO	Π_1	Π_1:7	0.0074	stable Π_1

Table 6: Loss-weight sensitivity on the three benchmarks where adaptive controllers are theoretically viable (SMS-Spam excluded as residual-bounded). We rerun the deployable $\Pi_0/\Pi_1/\Pi_2$ family pool under strict 5-fold-by-5-seed CV for seven local perturbations of the canonical weights $(1, 1, 0.05)$. The HallusionBench Π_1/Π_2 boundary is itself a quantitative diagnostic: under the canonical weights Π_2 edges Π_1 by only ≈ 3 seed-sd, and two perturbations swap the winner. A-OKVQA and FOLIO are fully stable across the seven variants, with FOLIO's 0.0074 minimum margin reflecting its 4.5 seed-sd canonical Π_1 gap.

OKVQA and FOLIO and flips it on HallusionBench in two of seven variants. HallusionBench's class call is therefore weight-dependent within the range we swept, and should be treated as a boundary case.

E.3 Automated class selection

Automated inner-CV class selection is near-best on the two Bernstein-viable benchmarks and fails on FOLIO when the regime-theoretic class restriction is withheld. Strict nested auto-pick loss differences relative to `always_direct` are $\Delta \in \{-0.183, -0.028, -0.003\}$ on {HallusionBench, A-OKVQA, FOLIO}. On HallusionBench the auto-pick lands within seed noise of the per-class Π_2 winner ($\Delta = -0.183$ vs. per-class Π_2 winner $\Delta = -0.188$, 0.7 joint seed-sd), and A-OKVQA remains close but more borderline ($\Delta = -0.028$ vs. per-class Π_2 winner $\Delta = -0.033$, 1.4 joint seed-sd). On FOLIO, the auto-pick reaches only $\Delta = -0.003$, materially below the per-class Π_1 winner at $\Delta = -0.033$ (a 0.030 absolute-loss gap, or 1.7 joint seed-sd when auto-pick variability is included) because at n=203 the unrestricted inner-CV pool occasionally lands on Π_2 families whose sign Theorem 2 explicitly does not certify. This is exactly the failure mode Theorem 4 (ii) predicts when $n < n_{\min}^{\mathrm{Bern}}$: empirical CV alone cannot reliably separate Π_1 from Π_2, and the regime diagnostic is what supplies the missing class restriction. The same per-fold class shares ($19/25$ Π_1, $4/25$ Π_2, $2/25$ Π_0) make this concrete: 4 outer-folds pick a Π_2 family at coverage q where Theorem 2 does not certify the sign, and these 4 folds are what drag the auto-pick mean from the per-class Π_1 winner to the much smaller -0.003. Restricting the candidate pool with the regime diagnostic ($\beta > 0$, $n < n_{\min}^{\mathrm{Bern}} \Rightarrow$ exclude Π_2 per Thm 4 (ii)) forces those 4 folds to be reselected from $\Pi_0 \cup \Pi_1$. The prospective selector of Section 3.1 does exactly this and reaches $\Delta = -0.0227$, against -0.003 here. ***HallusionBench*** picks Π_2 HGBC on $20/25$ outer-folds (HGBC-md4 on $12/25$, HGBC-md3 on $8/25$) and Π_1 KMeans on the remaining $5/25$, consistent with the per-class Π_2 winner in Table 7. ***FOLIO*** picks Π_1 KMeans on $19/25$ outer-folds (KM-K=6 on $6/25$, KM-K=8 on $5/25$, KM-K=4 and K=5 on $4/25$ each), the Π_2 family Selective-C=0.3 on $4/25$, and Π_0 `always_direct` on $2/25$; this dominant Π_1 class share matches the deployable per-class winner. ***A-OKVQA*** is the one benchmark where the inner-CV class selection is ambiguous: $10/25$ outer-folds pick the Π_1 winner KMeans-K=4 while the remaining $15/25$ are distributed across three Π_2 families (Selective C=0.3 on $10/25$, HGBC-md3 on $4/25$, HGBC-md4 on $1/25$), giving a 10:15 Π_1:Π_2 class split with the resulting auto-pick loss falling between the two per-class winners. This is precisely the high-signal Π_1/Π_2 boundary case predicted by Theorem 4 item (iii) when the Π_1 and Π_2 ceilings are commensurate; the per-class decomposition on outer-train remains a more informative summary than the inner-CV argmin in this regime. On TextVQA-OCR, the per-class decomposition selects Π_3 as the deployable top-rung winner because the OCR channel is available at prediction time. If the non-deployable A-OKVQA gold-rationale oracle is included in the candidate pool, inner-CV also selects it; we report that only as the oracle experiment of Appendix C.4, not as the deployable auto-pick result.

E.4 DIAGNOSTICS FOR THE FOUR CORE BENCHMARKS

Figure 7 plots the ceilings and the viability ratio for all ten benchmarks. Table 8 lists the underlying quantities numerically for the core four, including the intermediate $n\beta^2$ and α_{emp} that the figure does not show.

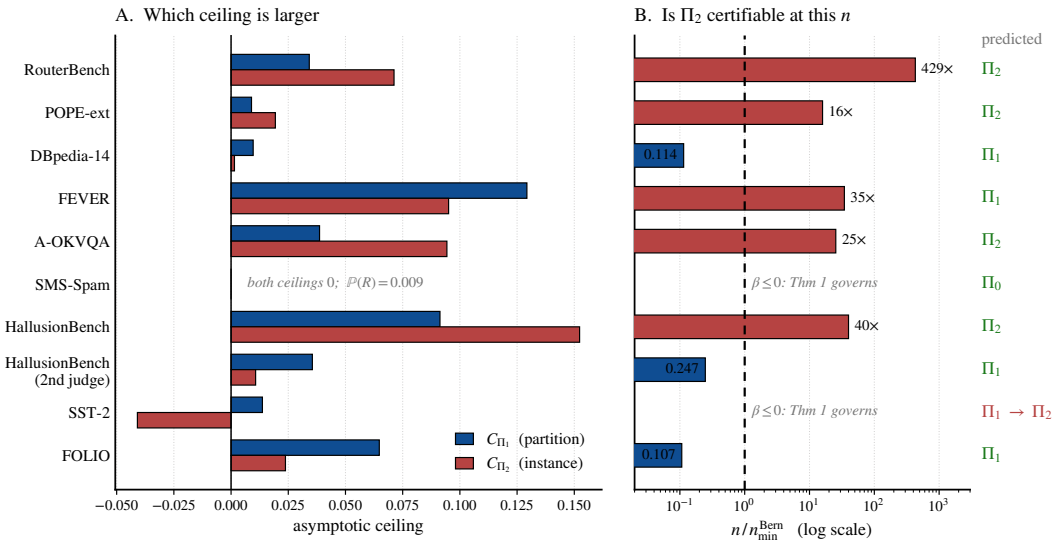

Figure 7: The two quantities that decide the class, across all ten benchmarks, ordered by n. **A**: the asymptotic ceilings C_{Π_1} and C_{Π_2}, whose comparison resolves branch (iii) of Theorem 4. **B**: the viability ratio $n/n_{\min}^{\text{Bern}}$ on a log scale, whose position relative to 1 resolves branch (ii); benchmarks with $\beta \leq 0$ have no threshold and fall to Theorem 1 instead. The predicted class on the right follows from the two panels alone. Nine of ten match the empirical winner. The exception, SST-2, is visible in Panel A as the one negative C_{Π_2}; Section 3.1 takes up what that costs.

Benchmark	Π_0 best↓	Π_1 best↓	Π_2 best↓	predicted
SMS-Spam (n=1114)	$\mathbf{0.0590 \pm 0.0000}$	$\mathbf{0.0590 \pm 0.0000}$ †	0.0591 ± 0.0000	Π_0
HallusionBench (n=920)	1.0000 ± 0.0000	0.9093 ± 0.0025	$\mathbf{0.8970 \pm 0.0039}$	Π_2
A-OKVQA (n=1145)	0.4138 ± 0.0000	0.3902 ± 0.0085	$\mathbf{0.3805 \pm 0.0032}$	Π_2
FOLIO (n=203)	0.7520 ± 0.0000	$\mathbf{0.7195 \pm 0.0073}$	0.7546 ± 0.0086	Π_1

Table 7: Strict nested 5-fold-by-5-seed CV per-class loss on the four core benchmarks (lower is better; values are mean ± seed-sd, fixed policies have zero seed variance). Bold marks the empirical winner within the canonical family pool fixed in advance (Appendix B.3); on A-OKVQA a CART realization of Π_1 admitted only as an ablation reaches a lower loss still (Appendix E.2); the rightmost column gives the regime-theoretic predicted winner from Theorem 2 and Theorem 4 (full predictions with diagnostics in Table 8). †: on SMS-Spam, Π_1 collapses to a_{direct} under CV, so $\Pi_1 = \Pi_0$ numerically; Π_0 wins, though Π_2 trails by only 0.0001 at the displayed precision. Per-benchmark mechanism (including `fair_fixed_train` on HallusionBench and the uncertified Π_2 outcome on FOLIO) is unpacked in the body.

E.5 ROUTING BASELINES: PER-BENCHMARK NUMBERS

We ran three canonical routing designs under the identical protocol, on the same loss matrices and the same feature block: a FrugalGPT-style cost-ordered cascade (Chen et al., 2023), a RouteLLM-style binary preference router (Ong et al., 2024), and a confidence-threshold deferral rule (Jitkrittum et al., 2023). All three are instance-level by construction and therefore Π_2-class; Table 9 reports them per benchmark.

The pattern is the one the theory predicts rather than a uniform advantage. Where the diagnostics select Π_2 the comparison is within-class, and our margin over the best baseline is 0.0147 on RouterBench,

Benchmark	n	α_{emp}	β	$n\beta^2$	$n_{\min}^{\text{Bern}}$	C_{Π_1}	C_{Π_2}	predicted class
SMS-Spam	1114	0.991	$\leq 0^{\dagger}$	—	—	0.000	—	Π_0 (residual-bounded; Thm 1)
HallusionBench	920	0.722	0.469	202.4	23	0.10	**0.15**	Π_2 (high-signal, $n \gg n_{\min}$; Thm 2)
A-OKVQA	1145	0.874	0.246	69.3	45	0.039	**0.094**	Π_2 (high-signal, Π_1/Π_2 boundary)
FOLIO	203	0.687	0.053	0.57	1898	**0.065**	$0.024^{\ddagger}$	Π_1 ($n < n_{\min}$, Π_2 uncertified; Thm 4 (ii))

Table 8: Theory-instantiated quantities from Theorem 2 and Theorem 4. $n_{\min}^{\text{Bern}}$ is computed from the leading-order formula $\lceil 2\alpha(1-\alpha)\log(2/\delta)/(q\beta^2)\rceil$ at δ=0.05, q=0.3, using the cross-validated AUC margin as the local tail-margin proxy. $C_{\Pi_1} = \max_{\mathcal{G}\in\mathfrak{G}} \sum_g p_g \gamma_g$ over the candidate KMeans family; $C_{\Pi_2} = q^\star \beta (L_w - L_r)$. Bold marks the diagnostic favoring the predicted adaptive class: for HallusionBench and A-OKVQA both classes are supportable and the larger ceiling indicates the Thm 4 (iii) winner; for FOLIO Π_2 is uncertified ($n < n_{\min}^{\text{Bern}}$, Thm 4 (ii)) so C_{Π_1} is bolded as the supportable winner. †SMS-Spam: $\beta \leq 0$, Lemma 1 rules out an instance-level rejector; $\mathbb{P}(R)$=0.009, oracle Π_2 gain =0.003. ‡FOLIO: C_{Π_2} is the asymptotic ceiling but Π_2's sign is uncertified at n=203 $< n_{\min}^{\text{Bern}}$, so the empirical winner falls to Π_1. Diagnostics are reported post-hoc on outer-test concatenations; the strict-CV selection rule uses only inner-fold quantities (Theorem 5; auto-pick analysis in Appendix E.3).

Benchmark	cascade	binary router	conf. deferral	selected class (ours)	class
RouterBench	0.5160 ± 0.0003	0.5304 ± 0.0005	0.5525 ± 0.0002	$\mathbf{0.5013 \pm 0.0003}$	Π_2
HallusionBench	1.0733 ± 0.0028	1.0650 ± 0.0054	0.9197 ± 0.0045	$\mathbf{0.8970 \pm 0.0039}$	Π_2
A-OKVQA	0.3809 ± 0.0034	0.3857 ± 0.0047	0.3825 ± 0.0022	$\mathbf{0.3805 \pm 0.0032}$	Π_2
FOLIO	0.7716 ± 0.0264	0.7622 ± 0.0195	0.7511 ± 0.0162	$\mathbf{0.7195 \pm 0.0073}$	Π_1
SMS-Spam	0.0598 ± 0.0007	0.0615 ± 0.0008	$\mathbf{0.0590 \pm 0.0000}$	$\mathbf{0.0590 \pm 0.0000}$	Π_0

Table 9: Three canonical routing designs under the identical strict nested 5-fold-by-5-seed protocol, against the best family inside the class the diagnostics select. All three baselines are Π_2-class by construction. Where the selected class is also Π_2 the comparison is within-class and the margins are small (A-OKVQA: 0.0004, within noise). The separation appears where the selected class is not Π_2: on FOLIO the best baseline trails by 0.0316, and on SMS-Spam the cascade and the binary router are both worse than doing nothing.

0.0227 on HallusionBench and 0.0004 on A-OKVQA—the last within seed noise. Where the diagnostics select a class other than Π_2 the gap opens: on FOLIO, which is variance-bounded, all three instance-level baselines trail the Π_1 router, the best of them (confidence deferral, 0.7511) by 0.0316; and on residual-bounded SMS-Spam the cascade and the binary router both increase loss relative to always-direct. Instance-level routing loses ground in the regimes the diagnostics single out, and holds its own elsewhere.

E.6 Direct tail-margin diagnostic

Section 3.1 states the summary; Table 10 gives the per-q measurements. $\tau(q) := \widehat{\mu}_w(q) - \alpha_{\min}$ is estimated on the held-out split against the fallback the tail is rejected to, and "agrees" records whether $\operatorname{sign}\tau(q) = \operatorname{sign}\beta$, which is the local tail-margin convention of Lemma 1 stated as a checkable condition rather than assumed.

Benchmark	β	$\tau(0.2)$	$\tau(0.3)$	$\tau(0.4)$	agrees
HallusionBench	$+0.469$	$+0.312$	$+0.265$	$+0.252$	3/3
A-OKVQA	$+0.246$	-0.017	-0.162	-0.248	0/3
FOLIO	$+0.053$	-0.033	$+0.033$	$+0.046$	2/3
SMS-Spam	-0.294	-0.923	-0.932	-0.937	3/3

Table 10: Direct bottom-q tail margin against the AUC-level proxy β. The convention holds outright on HallusionBench and SMS-Spam, holds at the operating q on FOLIO, and fails on A-OKVQA, where the deployed controller is four-action and the binary rejector that τ models is not the object actually fitted.

The A-OKVQA row is the informative one. Rejecting the bottom 30% wholesale to abstain increases loss by 0.033, which is what $\tau(0.3) = -0.162$ records; the deployed four-action controller instead reduces loss by 0.035 ± 0.003 on that same tail, allocating it (seed-averaged) as roughly 147 of 343 to defer, 51 to retrieve, 44 to abstain, and the remainder left on direct. The gap between these two numbers is the multi-action slack that τ cannot see, and it is the same slack that produces the SST-2 miss of Section 3.1, where it is large enough to overturn a negative β.

Substituting τ for β as the class gate is worse by one benchmark and errs on different ones: on the six benchmarks for which we computed both, gating on β names the empirical winner on five (missing SST-2) and gating on τ on four (missing FOLIO and DBpedia-14, on both of which the tail-derived ceiling exceeds C_{Π_1} and predicts Π_2 where Π_1 wins). We therefore keep β, whose error is one-sided in the safe direction—it understates instance-level gain and so errs toward the coarser class—and record τ alongside it.

E.7 Equal-information control for Π_3

Table 11 compares classes at the feature access Definition 1 assigns them: Π_3 gates on the OCR channel and the lower rungs do not see it. That comparison prices the channel and the gating form together. To price them separately we re-ran every class with the OCR block appended to its feature input (Table 12), so that a learned Π_2 controller and a Π_1 router have exactly the information the gate uses. Both tables come from one run over the full family pool at seeds $\{42, 73, 104, 135, 166\}$ on the TextVQA v0.5.1 release with Rosetta OCR v0.2, so the two halves share a seed set, a feature pipeline and a data snapshot.

TextVQA-OCR validation (n=5000)	best family	answer loss↓	Δ vs Π_3
Π_0	`always_ocr`	0.8449 ± 0.0000	$+0.0173$
Π_1	CART-`d4`	0.8410 ± 0.0017	$+0.0134$
Π_2	logistic C=0.3	0.8577 ± 0.0011	$+0.0301$
Π_3	OCR gate + fallback	$\mathbf{0.8276 \pm 0.0012}$	$\mathbf{0}$

Table 11: Deployable Π_3 on TextVQA-OCR, with the lower rungs restricted to the scalar non-OCR block as Definition 1 allows. Π_2 underperforms Π_0 because non-OCR features cannot represent OCR-side information; at n=5000, far above $n_{\min}$, this is a question of what the features encode. Π_3's 0.0134 margin over best-lower Π_1 is 6.4 joint seed-sd. The comparison prices the channel together with the gate; Appendix E.7 separates them.

TextVQA-OCR (n=5000)	family	answer loss↓	information
Π_0	`always_ocr`	0.8449 ± 0.0000	no OCR
Π_1	CART-`d4`	0.8410 ± 0.0017	no OCR
Π_2	logistic C=0.3	0.8577 ± 0.0011	no OCR
Π_2	HGBC-`d4` + OCR	0.8850 ± 0.0004	with OCR
Π_2	logistic C=0.3 + OCR	0.8471 ± 0.0017	with OCR
Π_1	CART-`d4` + OCR	0.8337 ± 0.0025	with OCR
Π_1	CART-`d3` + OCR	0.8337 ± 0.0019	with OCR
Π_3	OCR gate + fallback	$\mathbf{0.8276 \pm 0.0012}$	with OCR

Table 12: Every class given the OCR block, from the same run and the same seeds as Table 11. The channel is worth 0.0073 to the best lower class ($0.8410 \rightarrow 0.8337$); the gating form, handed that same channel, is worth a further 0.0060, or 2.6 joint seed-sd. Handing the channel to a learner instead of gating on it recovers less of it at every capacity we tried: the logistic plug-in reaches 0.8471 and the boosted-tree router 0.8850.

Both quantities clear seed noise: the channel is worth 0.0073 to the best lower class, and the gate a further 0.0060 at 2.6 joint seed-sd over the best class handed that same channel. The capacity ladder is the informative part. A logistic plug-in given the OCR block reaches 0.8471 and a boosted-tree router 0.8850, both short of the partition router at 0.8337 and of the gate at 0.8276, so the obstacle is not that the learners lack capacity but that estimating how to use the channel costs more at this n than reading it deterministically.

E.8 FOLIO WITH TEXT-SPECIALISED BACKBONES

FOLIO is a text-only logical-reasoning benchmark, and the core results obtain its three actions from Qwen2.5-VL run in text-only mode, which raises the possibility that its variance-bounded diagnosis is an artifact of using a vision-language backbone on text. We therefore regenerated all three actions with text-specialised backbones—Qwen2.5-3B-Instruct for direct and retrieve, Qwen2.5-7B-Instruct for defer—and rebuilt the loss matrix and the 12-dimensional feature block from those generations before re-running the identical pipeline (Table 13).

FOLIO (n=203)	Qwen2.5-VL (main text)	Qwen2.5-Instruct
Π_0 best	0.7520 ± 0.0000	0.7874 ± 0.0000
Π_1 best	$\mathbf{0.7195 \pm 0.0073}$	$\mathbf{0.7423 \pm 0.0253}$
Π_2 best	0.7546 ± 0.0086	0.7656 ± 0.0148
predicted class	Π_1	Π_1
empirical winner	Π_1	Π_1
selector class share	Π_1 25/25	Π_1 25/25

Table 13: FOLIO under the original vision-language backbones and under text-specialised backbones of the same sizes. Absolute losses shift and the family ordering inside Π_1 changes (K=6 to K=4), but the regime tag, the predicted class, the empirical winner and the prospective selector's per-fold class choice are unchanged. The Π_1–Π_2 gap is 0.79 joint seed-sd under the text backbones and therefore not resolved above seed noise; what the ablation establishes is that the variance-bounded diagnosis is not a backbone artifact, not that Π_1 strictly separates from Π_2 here.