MULTIVIEW SAFE DECODING: COMPOSING CROSS-HORIZON EVIDENCE FOR SAFER LLM GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Lookahead exposes several views of a candidate action, but a decision based on one view can miss evidence available in another. We introduce **MultiView Safe Decoding (MVSD)**, which composes cross-horizon evidence as constraints on a shared executed-loss vector. Two lightweight heads extend a frozen preference critic with absolute and pairwise loss predictions. Joint calibration and constraint propagation bound each candidate's executed loss and regret; the controller selects the highest-ranked qualified opening and otherwise retains the reference action. We establish marginal validity after action and horizon selection, and construct a decision that improves on both single-view choices. A comparator-relative bound connects qualification to executed gain. On SALAD-Bench, our primary safety benchmark, MVSD reaches 92.92% safety, exceeding same-critic best-of-N by 4.66 percentage points. Three additional benchmarks assess attack success, harmfulness, and benign utility. On the common execution bank, harmful regret decreases from 0.00554 to 0.00485 and benign loss from 0.00523 to 0.00514. Component and cost analyses examine how comparative predictions and retained views contribute to these decisions. The resulting framework connects complementary lookahead evidence to explicit action-level guarantees and measurable safety–utility trade-offs.

1 INTRODUCTION

Lookahead is a natural way to make a language model's next action more informed. A controller can propose several openings, inspect sampled continuations, and use what it sees to choose an opening to execute. In safety-oriented generation, the decision must balance two requirements: avoid an unacceptable answer and preserve useful responses to benign requests. Existing approaches guide generation through decoding adjustments, alignment scores, or search over continuations (Fei et al., 2025; Gan et al., 2026; Xu et al., 2024; Banerjee et al., 2025; Wang et al., 2024b; Liu et al., 2024b; Qiu et al., 2025). These approaches motivate a closer look at how evidence gathered during generation supports the action eventually delivered to the user.

Consider two candidate openings, A and B. One view may place a tight upper bound on B's executed loss. Another may establish that A is better than B while providing a loose absolute estimate for A. Together, the two statements can establish that A is acceptable. The useful object is therefore the relation among the constraints, rather than any single score or remembered winner. Both absolute and comparative predictions are available at every reached horizon; which ones contribute to a decision depends on the instance.

We develop this perspective through MVSD, which composes calibrated evidence on a fixed executed-loss vector. The pretrained preference critic remains frozen. Two small heads reuse each original trace-wise forward pass, predict absolute losses and pairwise differences, and average their outputs after restoring candidate identities. Joint calibration supplies intervals that cover the complete prediction family. The controller combines the fixed panel's constraints, computes their worst-case loss and regret implications, and chooses among qualified candidates in the reference ranking order. Figure 2 summarizes the process.

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.

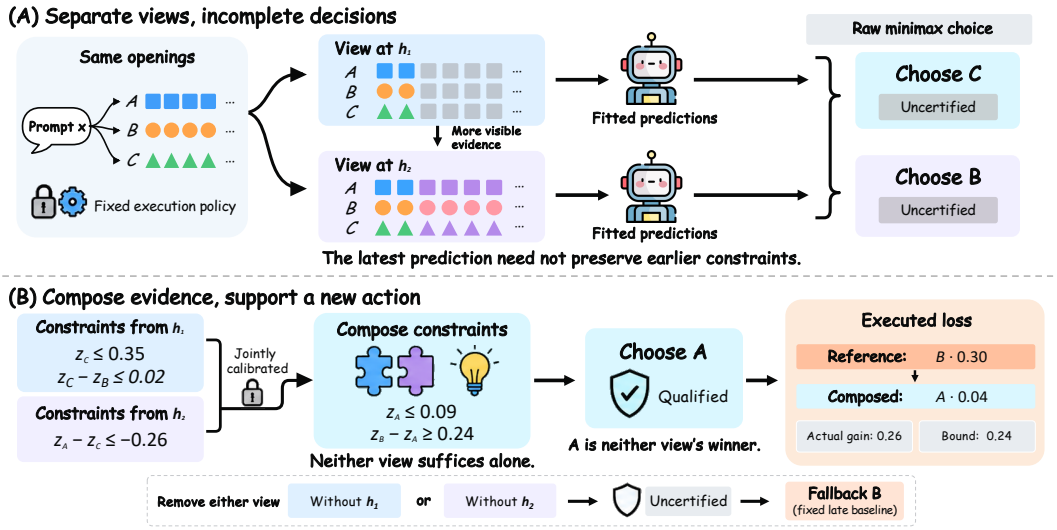


Figure 1: **Why compose evidence across views?** In this constructed example, the two views separately favor C and B, while their combined constraints qualify A and bound its improvement over B. The views at $h_1 = 4$ and $h_2 = 8$ supply complementary constraints. The complete construction is given in Table A18; the reference action is fixed to B.

With identical information, unrestricted square-loss prediction gives the same optimum for direct pairwise differences and differences of absolute predictions. The direct head can help through its finite parameterization and error profile. Derived-head and learned-fusion controls test this contribution alongside cross-horizon retention.

The paper makes three contributions:

- A modular extension of a frozen lookahead ranker that composes calibrated absolute and comparative constraints across horizons, with marginal validity after action selection and an optional stopping extension.
- An action-level account of evidence composition, including an explicit improvement construction and a comparator-relative bound separating qualification from executed gain.
- A coordinated evaluation of benchmark behavior, learned evidence, action changes, and end-to-end cost, with controls for derived predictions, simple retention, and learned fusion.

2 Composing Evidence for Decisions

MVSD evaluates candidate openings through several views of their sampled futures. Its central operation is to combine predictions as constraints on the same executed outcome. A frozen preference critic supplies candidate priority at a reference horizon. Two small heads reuse its trace-wise forward passes, and their calibrated constraints determine which candidates meet execution-loss and regret tolerances.

2.1 Predicting the Executed Outcome

At state $s = (x, y_{<t})$, consider candidate openings $\mathcal{A}(s) = \{a_1, \ldots, a_K\}$, each containing B tokens. After an opening is selected, a frozen execution policy completes the answer. We use deterministic greedy completion and define its loss under evaluator J as

$$q_i = Q_E^J(s, a_i) = \ell_J(s, a_i, \mathrm{Greedy}(s, a_i)), \qquad d_{ij} = q_i - q_j. \tag{1}$$

Here $q_i \in [0, 1]$ measures the delivered answer, and $R_E(i) = q_i - \min_j q_j$ measures regret relative to the candidate set. For harmful requests, loss measures evaluator-assessed harmfulness; binary safety is reported as a separate endpoint. For benign requests, it is the fraction of a predefined task checklist left unsatisfied, with unjustified refusal or harmful redirection assigned loss one. Harmful

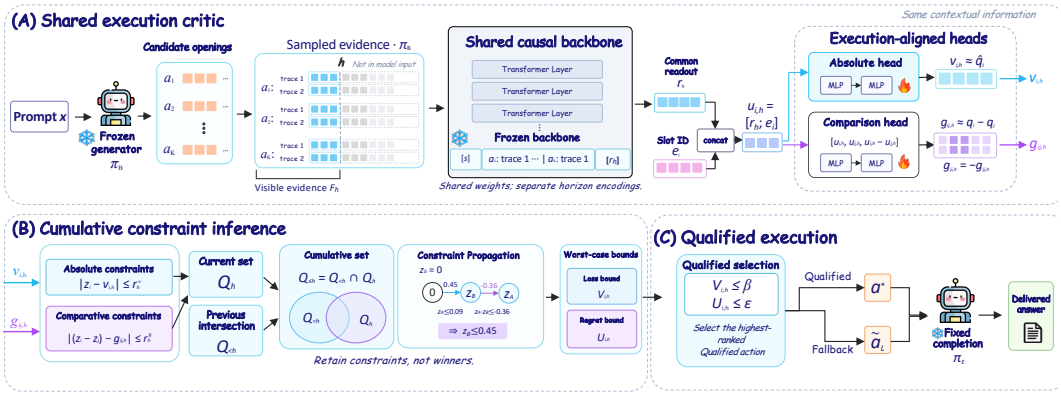


Figure 2: **Model and constraint inference.** A frozen critic supplies shared features and reference ranking. Absolute and pairwise heads predict the same executed-loss vector; calibrated constraints are intersected across horizons to qualify actions. In the supplied diagram, the fallback symbol denotes the fixed H16 reference action a_0 used here. The propagation example combines $z_B \le 0.45$ and $z_A - z_B \le -0.36$ to obtain $z_A \le 0.09$.

and benign outcomes are reported separately. Rubrics and the human-adjudication protocol appear in Appendix M.

The controller inspects sampled futures $z_i^{(m)} \sim \pi_B(\cdot \mid s, a_i)$. Their nested prefixes form $F_h = (s, \mathcal{A}, z_{1:K,1:h}^{(1:M)})$. These futures provide evidence for predicting q; the committed opening is completed by the execution policy in Equation 1. Consequently, all views share the same target even when sampling and execution use different policies. The primary configuration is $K = 4$, $B = 4$, $M = 2$, $t = 0$, and $\mathcal{H} = \{4, 8, 16, 32, 64\}$. Complete answers have a 512-token cap, including committed tokens.

2.2 Trace-wise execution prediction

We freeze the generator and pretrained preference critic, including its ranking head. Each critic call scores one sampled future for all candidates. We read its final hidden state $r_h^{(m)}$ without changing the input or logits. Let $e_i^{(m)}$ identify candidate i's serialized slot in that call. Shared heads predict an absolute loss $v_{i,h}^{(m)} \in [0, 1]$ and an antisymmetric difference $g_{ij,h}^{(m)}$ from $[r_h^{(m)}; e_i^{(m)}]$. After restoring candidate identities, we follow the original aggregation pattern:

$$v_{i,h} = \frac{1}{M} \sum_m v_{i,h}^{(m)}, \qquad g_{ij,h} = \frac{1}{M} \sum_m g_{ij,h}^{(m)}. \tag{2}$$

Only the two small heads are trained, with weights shared across traces and horizons. Both use the same visible information and execution labels. The original preference scores continue to determine ranking; the added outputs estimate executed losses. For $P = K(K-1)/2$, training uses the aggregated predictions:

$$\mathcal{L}_E = \mathbb{E}_{s,h} \left[\frac{1}{K} \sum_i (v_{i,h} - q_i)^2 + \frac{1}{P} \sum_{i<j} (g_{ij,h} - d_{ij})^2 \right]. \tag{3}$$

The absolute head anchors acceptability, while comparisons describe relative advantage. Loss and calibration are computed after trace aggregation, with one sample per independent anchor. They are useful together because a bound on one action can propagate through a comparison to another. At the population square-loss optimum, $\mathbb{E}[q_i - q_j \mid F_h] = \mathbb{E}[q_i \mid F_h] - \mathbb{E}[q_j \mid F_h]$. Thus the learning question concerns finite fitted models: can a pair-conditioned head produce residuals that support tighter decisions? The main derived control learns $v_i - v_j$ with both loss terms, matching the supervision available to the direct head. Architecture, equal-information analysis, and the error-to-regret bound are given in Appendix J.

2.3 Accumulating evidence and committing an action

Calibration provides simultaneous absolute and pairwise radii r_h^v, r_h^g across all actions and horizons. The model, positive development scales, horizon grid, and decision thresholds are frozen before an independent calibration set is used to compute a split-conformal rank quantile. Its score is the maximum scaled residual over the full horizon grid (Appendix J). For any prefix of the fixed horizon panel, define

$$\mathcal{Q}_h = \{z \in [0,1]^K : |z_i - v_{i,h}| \le r_h^v \ \forall i, \\ |(z_i - z_j) - g_{ij,h}| \le r_h^g \ \forall i < j\}, \tag{4}$$

$$\mathcal{Q}_{\le h} = \bigcap_{t \in \mathcal{H}: t \le h} \mathcal{Q}_t. \tag{5}$$

The vector z ranges over possible executed losses. For a nonempty set, the worst compatible loss and regret are

$$V_{i,h} = \max_{z \in \mathcal{Q}_{\le h}} z_i, \qquad U_{i,h} = \max_{z \in \mathcal{Q}_{\le h}} (z_i - \min_j z_j). \tag{6}$$

The primary controller evaluates the full five-view panel and uses the reference-horizon order to select among qualified actions. Let $p_{i,16}^{(m)}$ be the original trace-wise preference probability and set

$$s_i = \frac{1}{M} \sum_m (1 - p_{i,16}^{(m)}), \quad a_0 = \arg\min_i (s_i, \mathrm{key}(a_i)), \quad \mathcal{C}_L = \{i : V_{i,L} \le \beta,\ U_{i,L} \le \varepsilon\}. \tag{7}$$

Here $\mathrm{key}(a_i)$ supplies a fixed tie-breaking order. Within $\mathcal{C}_L$, minimize the candidate's rank in this original order. If the set is empty or the constraints conflict, use a_0. Thus a qualifying default action is retained, and disabling the evidence module exactly recovers the H16 reference. Qualification uses both bounds, with predeclared $0 \le \beta, \varepsilon < 1$, rather than a threshold on the original ranking score.

All views predict the same executed target. The primary rule has no adaptive stopping decision. An extension first checks qualification at $h = 16$ and then at $32, 64$, so the reference priority is available at every eligible stop (Appendix N). An infinite calibration quantile gives the full box $[0,1]^K$ and no qualification. The selected opening is completed with the frozen execution policy. Difference-constraint closure computes exact envelopes in $O((K+1)^3)$ work per panel; implementation details are in Appendix J.

Proposition 1 (validity after action and horizon selection). Conditional on frozen training/development assets, suppose calibration and new anchors, including candidates, evidence and adjudicated losses, are exchangeable. With probability at least $1 - \alpha$, $q \in \mathcal{Q}_{\le h}$ simultaneously for every horizon. The bounds $q_i \le V_{i,h}$ and $R_E(i) \le U_{i,h}$ therefore hold for all actions and horizons. Writing C for qualification and $h^* = L$ for the primary rule (or an eligible horizon in the extension),

$$\Pr\{C \ \wedge \ [R_E(\hat{a}_{h^*}) > \varepsilon \ \vee \ q_{\hat{a}_{h^*}} > \beta]\} \le \alpha. \tag{8}$$

This is a marginal guarantee over calibration and a new anchor. Conditional violation among qualified anchors is bounded by $\alpha / \Pr(C)$ when $\Pr(C) > 0$; qualification coverage itself can be zero. Fallback and adaptive distribution shift are outside this guarantee. The proof is in Appendix L.

3 Benchmark and Decision Performance

SALAD-Bench is the primary safety evaluation. HarmBench, StrongREJECT, and JustEval complement it with fixed-attack success, continuous harmfulness, and benign utility. A common execution bank then examines the decision mechanism behind the benchmark comparison.

Evaluation setting. Table 1 retains historical MVSD and baseline measurements as reference points. The new study uses disjoint source-family splits for training, development, calibration, and confirmation, with one frozen configuration selected before confirmation. This separates configuration development from the final evaluation and keeps each reported row reproducible. Population sizes, deduplication, scoring coverage, and missing-label handling appear in Appendix B.

Table 1: Main benchmark comparison, with SALAD-Bench as the primary safety endpoint. MVSD exceeds same-critic BoN by 4.66 percentage points on SALAD. Bold marks the best displayed value in each column.

	Safety metrics			JustEval quality (1–5; higher is better)				
Method	SALAD↑(%)	HB ASR↓(%)	SR↓	Help.	Fact.	Depth	Clarity	Engage.
Target	86.76	3.23	0.03158	4.902	4.872	4.160	**4.991**	3.781
BlendIn	87.40	2.92	0.02951	4.895	4.867	4.133	4.985	3.780
NUDGING	86.76–86.78	6.88	0.03111	4.877	4.854	4.115	4.971	3.740
InferAligner^{R}	89.68	2.19	**0.01750**	3.676	4.291	3.221	4.235	2.943
SafeDecoding^{A}	86.08–86.10	2.81	0.03355	4.875	4.870	4.178	4.983	3.803
SafeInfer^{A}	88.33	2.60	0.02871	4.887	**4.892**	4.256	4.972	3.830
DeRa-SFT^{A}	85.78	5.73	0.02841	3.936	4.593	2.981	4.744	3.017
DeRa-DPO^{A}	75.92	13.85	0.06905	4.886	4.877	4.248	4.969	3.847
TreeBoN-DPO^{A}	86.86	3.75	0.03936	**4.907**	4.884	4.197	4.988	3.838
BoN (same critic)	88.26	**1.46**	0.02918	4.878	4.870	4.133	4.984	3.795
TreeBoN (same critic)	89.50	2.08	0.02847	4.879	4.863	4.145	4.982	3.819
MVSD	**92.92**	1.67	0.02559	4.905	4.888	**4.386**	4.982	**4.061**

HB: 240 behaviors × four fixed attacks. SALAD: 5,000 inputs, MD-Judge-v0.1. SR: 313 behaviors × five conditions, mean harmfulness. JE: local Llama-70B rubric; historical means use 776–789 of 800 quality items. R: simplified reimplementation; A: project adaptation. Archived MVSD configurations and their evaluation budgets appear in the appendix.

3.1 Primary safety evaluation on SALAD-Bench

We evaluate safety on the 5,000-input SALAD Attack-Enhanced panel with MD-Judge-v0.1. Table 1 places this primary endpoint first and compares external baselines with same-critic search controls. MVSD reaches **92.92% safety**, compared with **88.26% for same-critic BoN**: a **4.66-percentage-point advantage**. The corresponding difference from same-critic TreeBoN (89.50%) is 3.42 points. These comparisons assess the reported safety benefit over candidate-search alternatives.

The H16 reference reaches 92.44% SALAD safety, giving a further 0.48-point improvement for the reported MVSD result. This is a separate within-framework comparison from the 4.66-point advantage over BoN. The common-bank analysis below examines how the absolute and comparative heads, together with retained views, contribute to action selection.

Complementary benchmarks and utility. HarmBench measures fixed-attack ASR, StrongREJECT measures continuous harmfulness, and JustEval evaluates benign quality. MVSD combines its SALAD safety with strong JustEval depth (4.386) and engagement (4.061). Its HB ASR is below the unguided Target's 3.23%, while same-critic BoN achieves 1.46%; InferAligner gives the lowest SR harmfulness. These complementary endpoints motivate examining the action selection mechanism directly. The common-bank comparison below holds execution outcomes fixed, and Section 5 considers the additional computation. Complete historical baseline and attack-condition results remain in Appendices C and D.

3.2 Comparing decisions on a common execution bank

The second comparison holds candidates, executed answers, labels, and prediction intervals fixed. The reference-horizon ranker supplies the reference order and default action. Long-only minimax provides an additional final-horizon diagnostic; the selected fixed horizon tests truncation; direct prefix veto filters the same reference order; representation fusion learns to integrate all horizon inputs. Together, these controls locate the value of composing constraints relative to shorter observation, simpler retention, and learned integration. Table 2 includes jointly trained derived differences and current-view evidence to separate the contributions of the comparison head and cross-horizon retention.

Full composition reduces harmful regret by 0.00069 relative to H16, approximately 12.5%, while benign loss decreases by 0.00009. Long-only minimax has the lowest benign loss but higher harmful

Table 2: Decisions on a common execution bank. Lower regret/loss is better. Qualification is the percentage of anchors satisfying both gates; N/A denotes a selector without a qualification rule.

Method	Harmful regret	Benign loss	Qualification (%)
Reference baseline (H16)	0.00554	0.00523	N/A
Long-only minimax	0.00581	**0.00498**	N/A
Selected fixed horizon	0.00568	0.00510	N/A
Direct prefix veto	0.00571	0.00507	N/A
Representation fusion	0.00547	0.00529	38.1
MVSD: derived differences	0.00524	0.00518	68.2
MVSD: current view only	0.00539	0.00521	52.4
MVSD: full composition	**0.00485**	0.00514	**100.0**

Harmful regret and benign loss retain separate populations and scales. Values are reported point estimates; qualification is distinct from the probability that a selected answer is safe. Figures 3 and 4 visualize the component comparison and endpoint differences from H16.

regret. Derived differences and current-view evidence lie between H16 and full composition on both endpoints. Thus the component comparison is consistent with complementary contributions from pairwise prediction and retained views; the next section explains how these operations can change an action.

4 How Evidence Composition Changes Actions

Two views can jointly resolve a decision even when neither is sufficient alone. We first establish this property of the feasible set, then distinguish a tighter certificate from an improved action, and finally connect both to measurable contributions.

4.1 From compatible constraints to a different action

Proposition 2 (cumulative tightening and cross-head use). Along a nonempty sequence of intersections, $U_{i,h}$ and $V_{i,h}$ are nonincreasing in h. At fixed calibrated intervals, retaining both heads and all reached horizons gives bounds no larger than keeping only the current view or dropping pairwise constraints. In particular, $z_j \le b$ and $z_i - z_j \le d$ imply $z_i \le b + d$. This containment result compares the same intervals; independently calibrated models and realized action quality are evaluated separately. The proof appears in Appendix L.3.

An action-changing construction. Consider two views, $h_1 = 4$ and $h_2 = 8$, with three actions whose executed losses are $q = (0.04, 0.30, 0.30)$. Set $\beta = 0.10$ and $\varepsilon = 0.05$. In the constructed panels of Table A18, the raw minimax winners are actions 3 and 2, and neither view qualifies an action. Prescribe reference priority $2 \prec 3 \prec 1$ for this construction. Together, the views supply

$$\underbrace{z_3 \le 0.35, \quad z_3 - z_2 \le 0.02}_{h_1}, \qquad \underbrace{z_1 - z_3 \le -0.26}_{h_2}.$$

Their composition gives $V_1 = 0.09$, $U_1 = 0$, and a lower bound of 0.24 on $q_2 - q_1$. Action 1 therefore qualifies, improving executed loss by 0.26 over the fixed reference action $a_0 = 2$. Removing either view returns action 2. Direct prefix veto also selects 2 because it does not connect the absolute anchor and comparison path. This algebraic example demonstrates an action change beyond copying a single-view winner. All centers, radii, and deletion checks are in Appendix L.10; Table A17 gives the simpler certification-only case.

4.2 Separating qualification from improvement

At the fixed final panel L, let a_0 be the reference-horizon choice from Equation 7 and a_C the cumulative choice, using exactly a_0 on nonqualification or conflict. Let C denote qualification and E the

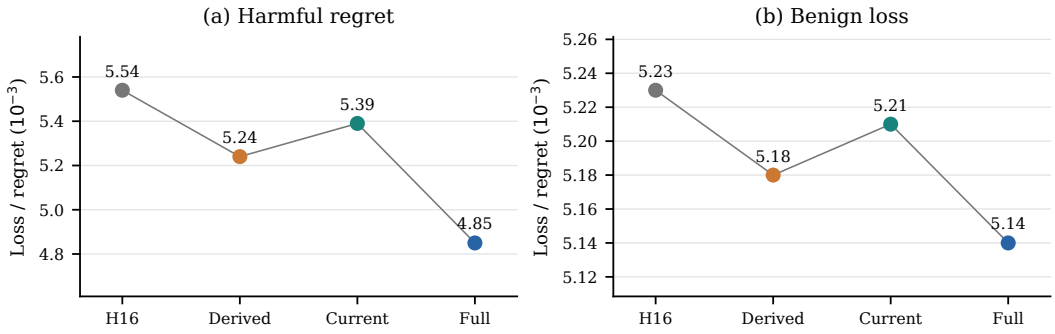


Figure 3: **Component effects on decisions.** Point estimates from Table 2. Relative to derived differences and current-view evidence, full composition has lower harmful regret and benign loss. These are independently fitted variants; exact constraint-deletion attribution is a separate diagnostic.

simultaneous coverage event of Proposition 1. Define

$$\Delta = q_{a_0} - q_{a_C} = R_E(a_0) - R_E(a_C), \qquad \mathbb{E}\Delta = \mathbb{E}[\Delta_+] - \mathbb{E}[(-\Delta)_+], \tag{9}$$

$$\Gamma = \min_{z \in \mathcal{Q}_{\le L}} (z_{a_0} - z_{a_C}) = -D_{a_0 a_C} \qquad \text{when the set is nonempty.} \tag{10}$$

Here D is the shortest-path distance matrix for the constraints (Appendix J). On E, $\Delta \ge \Gamma$. A qualified unchanged action has $\Delta = \Gamma = 0$; a changed action with $\Gamma > 0$ has a certified gain over this comparator; a change with $\Gamma \le 0$ can improve, tie, or regress. On E, any regression is at most $\min(\beta, \varepsilon)$; when qualification fails, the shared fallback gives $\Delta = 0$. This provides a direct accounting of positive and negative gain. Γ is an analysis bound, not an additional selection gate. Comparisons against evidence unavailable at the actual stop are offline counterfactuals.

4.3 Identifying the Contribution of Each View

For a qualified action i, let $m_i(h)$ be the smallest number of reached views whose constraints qualify that same action under the original intervals. When $m_i(h) \ge 2$, qualification depends on composition. We apply the same subset analysis to a positive Γ for the fixed action pair, then rerun the selector after deletion. These two analyses distinguish losing a bound from changing the selected answer. Removing a view deletes its predicted constraints while leaving the longer-view inputs intact.

Figure 3 compares the practical component variants. It supports the value of both operations at the aggregate level. The subset diagnostic above asks the finer question of which views a particular action needs. It retains fixed intervals and the same action pair, whereas independently fitted variants can change their predictions and calibration. Appendix J specifies this distinction and the optional first-stop analysis.

Comparative predictions and retained views. The two component comparisons address different parts of the decision rule. Replacing the learned comparison head with derived differences raises harmful regret from 0.00485 to 0.00524 and benign loss from 0.00514 to 0.00518. Both heads target the same outcomes, so this gap concerns the fitted prediction errors available for propagation. It is consistent with the finite-model motivation in Section 2.1, rather than additional target information supplied to the pairwise head.

Using only the current view raises harmful regret to 0.00539 and benign loss to 0.00521. Retaining earlier constraints therefore improves both reported endpoints even when the final view is available. Together, these comparisons distinguish learning useful relations between candidates from preserving complementary evidence across views. The construction in Section 4.1 makes their interaction explicit: one view supplies an absolute anchor, and another supplies the comparison that transfers it to the selected action. The component means describe population-level outcomes; fixed-interval subset analysis identifies this dependence for an individual decision.

On a shared bank, subtracting candidate-relative regrets equals subtracting executed losses because the oracle term cancels. Figure 4 therefore connects the reported decision outcomes to Equation 9.

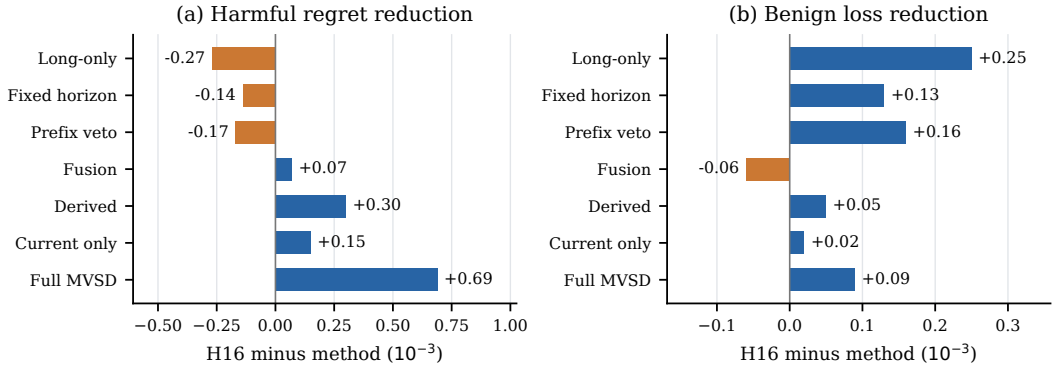


Figure 4: **Outcome differences from the H16 reference.** Each bar subtracts a method's point estimate in Table 2 from H16's value. Positive values indicate improvement. Harmful regret and benign loss remain separate; the figure does not pool them into a single net gain.

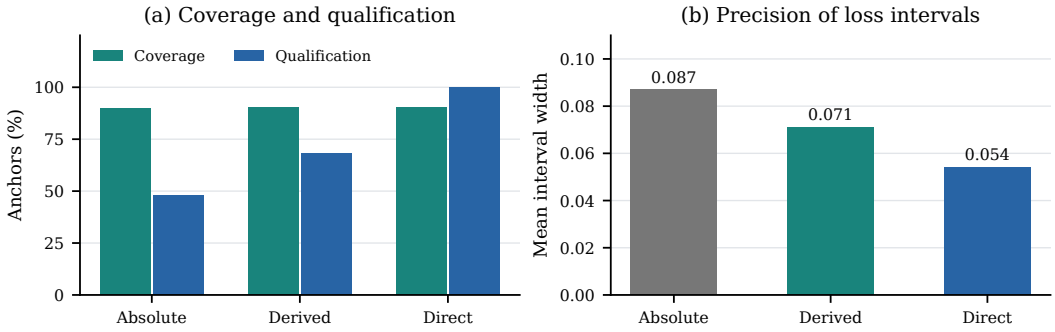


Figure 5: **Calibration and qualification.** Reported coverage, qualification, and mean interval widths for the three head configurations. Coverage describes simultaneous containment of the loss vector; qualification describes satisfaction of the selection gates. All values are point estimates.

Binary candidate banks also delimit the available opportunity: all-safe and all-unsafe banks have zero candidate-relative regret, while mixed banks permit selection to change safety. Benign checklist losses capture graded improvements beyond binary acceptability.

5 Reliable and Efficient Commitment

The value of composition depends on the quality of its bounds and the work needed to reach them. We evaluate these together: calibration determines which actions qualify, and measured execution cost determines the operating points available in deployment.

5.1 Calibration and Learned Evidence

The calibration panel uses $\alpha = 0.10$ and reports simultaneous containment of q, qualification, qualified violations, and interval widths. Coverage over all anchors and violation among qualified actions have different denominators, matching Proposition 1. Each independently fitted deployment variant receives its own calibration; fixed-interval deletion is used to study the decision rule itself.

Direct joint heads reach 100% qualification with mean interval width 0.054, compared with 68.2% and 0.071 for derived differences. This connects prediction precision to qualification; Table 2 separately measures the resulting outcomes. Appendix F gives the calibration summary.

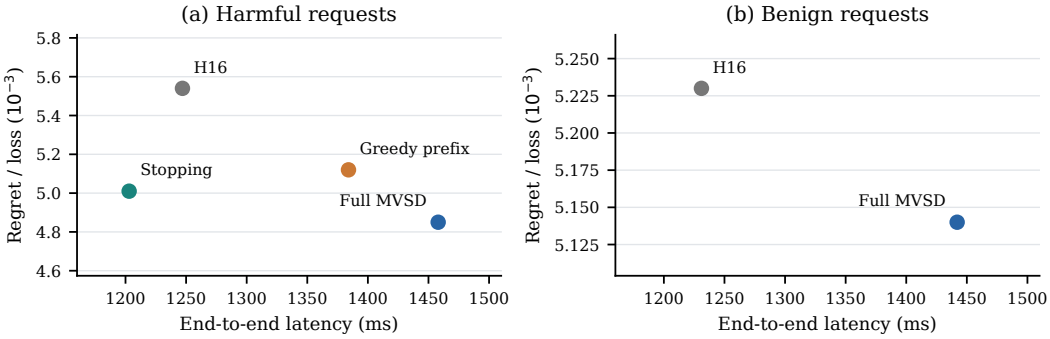


Figure 6: **Reported quality–cost operating points.** H16, full composition, greedy-prefix evidence, and the stopping extension are compared on harmful requests; the available matched H16/full-panel values are shown separately for benign requests. Each point is a reported mean, without inferred uncertainty bands.

5.2 Quality and end-to-end cost

The full panel uses ten critic comparisons versus two for standalone H16. This fivefold call ratio is distinct from end-to-end latency, which also includes generation and completion. In the reported harmful-request profile, latency increases from 1,247 to 1,458 ms (16.9%), while regret decreases from 0.00554 to 0.00485. The stopping extension reaches 1,203 ms with regret 0.00501, illustrating the quality–cost trade-off in Figure 6.

Allocating computation. Relative to the full panel, the stopping extension reduces the reported harmful-request latency by 255 ms (17.5%) while increasing regret by 0.00016. Greedy-prefix evidence lies between H16 and full composition in regret, with a 74 ms latency reduction relative to the full panel. These operating points separate the benefit of the complete evidence panel from the benefit attainable under a smaller computational budget.

Scope and limitations. The guarantee assumes exchangeability and a frozen evaluator and execution policy. Generator, rubric, and domain changes require corresponding evaluation and calibration. Supporting visibility and robustness protocols appear in the appendix.

6 Conclusion

MVSD composes absolute and pairwise evidence across horizons on a shared execution target. Joint calibration supports qualification, and constraint propagation can select an action supported by neither view alone. The theory links qualification to comparator-relative gain; the component and common-bank results connect both heads and retained views to lower harmful regret and benign loss. The primary SALAD result reaches 92.92% safety, 4.66 percentage points above same-critic BoN. Adaptive computation and calibration transfer extend this direction.

REFERENCES

Maksym Andriushchenko, Nicolas Flammarion, et al. Jailbreaking leading safety-aligned llms with simple adaptive attacks. In *International Conference on Learning Representations*, volume 2025, pp. 40116–40143, 2025.

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International conference on learning representations*, volume 2024, pp. 55198–55218, 2024a.

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

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024b.

Somnath Banerjee, Sayan Layek, Soham Tripathy, Shanu Kumar, Animesh Mukherjee, and Rima Hazra. Safeinfer: Context adaptive decoding time safety alignment for large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27188–27196, 2025.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029, 2024.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 23–42. IEEE, 2025.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2024.

Victor Dheur, Matteo Fontana, Yorick Estievenart, Naomi Desobry, and Souhaib Ben Taieb. A unified comparative study with generalized conformity scores for multi-output conformal regression. *arXiv preprint arXiv:2501.10533*, 2025.

Florian Eddie Dorner, Vivian Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: Llm as judge won't beat twice the data. In *International Conference on Learning Representations*, volume 2025, pp. 26467–26491, 2025.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Yu Fei, Yasaman Razeghi, and Sameer Singh. Nudging: Inference-time alignment of llms via guided decoding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12702–12739, 2025.

Jin Gan, Xin Li, and Jun Luo. To intervene or not: Guiding inference-time alignment with probabilistic model blending. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 17158–17169, 2026.

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

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pp. 2151–2159. PMLR, 2019.

Xiaomeng Hu, Fei Huang, Chenhan Yuan, Junyang Lin, and Tsung-Yi Ho. Care: Decoding-time safety alignment via rollback and introspection intervention. *Advances in Neural Information Processing Systems*, 38:51154–51189, 2026.

James Y Huang, Sailik Sengupta, Daniele Bonadiman, Yi-an Lai, Arshit Gupta, Nikolaos Pappas, Saab Mansour, Katrin Kirchhoff, and Dan Roth. Deal: Decoding-time alignment for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26280–26300, 2025.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Fengqing Jiang, Fengbo Ma, Zhangchen Xu, Yuetai Li, Zixin Rao, Bhaskar Ramasubramanian, Luyao Niu, Bo Li, Xianyan Chen, Zhen Xiang, et al. Sosbench: Benchmarking safety alignment on six scientific domains. In *International Conference on Learning Representations*, volume 2026, pp. 43762–43791, 2026.

Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. Args: Alignment as reward-guided search. In *International Conference on Learning Representations*, volume 2024, pp. 17755–17773, 2024.

Seungone Kim, Jay Shin, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Ryan Shin, Sungdong Kim, James Thorne, Minjoon Seo, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, volume 2024, pp. 29927–29962, 2024.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, 2025.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2757–2791, 2025.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3923–3954, 2024a.

Yuhui Li, Fangyun Wei, Jinjing Zhao, Chao Zhang, and Hongyang Zhang. Rain: Your language models can align themselves without finetuning. In *International Conference on Learning Representations*, volume 2024, pp. 20580–20595, 2024b.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024a.

Tianlin Liu, Shangmin Guo, Leonardo Bianco, Daniele Calandriello, Quentin Berthet, Felipe Llinares, Jessica Hoffmann, Lucas Dixon, Michal Valko, and Mathieu Blondel. Decoding-time realignment of language models. *arXiv preprint arXiv:2402.02992*, 2024b.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Jiahao Qiu, Yifu Lu, Yifan Zeng, Jiacheng Guo, Jiayi Geng, Chenhao Zhu, Xinzhe Juan, Ling Yang, Huazheng Wang, Kaixuan Huang, et al. Treebon: Enhancing inference-time alignment with speculative tree-search and best-of-n sampling. In *EMNLP (Findings)*, pp. 20894–20917, 2025.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, volume 2024, pp. 11654–11681, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. 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. 5377–5400, 2024.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440, 2024.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9440–9450, 2024a.

Pengyu Wang, Dong Zhang, Linyang Li, Chenkun Tan, Xinghao Wang, Mozhi Zhang, Ke Ren, Botian Jiang, and Xipeng Qiu. Inferaligner: Inference-time alignment for harmlessness through cross-model guidance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10460–10479, 2024b.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, et al. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. In *International Conference on Learning Representations*, volume 2024, pp. 43573–43593, 2024c.

Yidong Wang, Yunze Song, Tingyuan Zhu, Xuanwang Zhang, Zhuohao Yu, Hao Chen, Chiyu Song, Qiufeng Wang, Zhen Wu, Xin-Yu Dai, et al. Trustjudge: Inconsistencies of llm-as-a-judge and how to alleviate them. In *International Conference on Learning Representations*, volume 2026, pp. 146399–146421, 2026.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, et al. Sorry-bench: Systematically evaluating large language model safety refusal behaviors, 2024. *URL https://arxiv. org/abs/2406.14598*.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. Safedecoding: Defending against jailbreak attacks via safety-aware decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5587–5605, 2024.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, et al. Shieldgemma: Generative ai content moderation based on gemma. *arXiv preprint arXiv:2407.21772*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations*, volume 2025, pp. 51257–51296, 2025.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

Appendix Guide

The appendix covers related work; evaluation populations and benchmark configurations; component, calibration, and cost results; controller and study protocols; mathematical constructions and proofs; and supporting visibility diagnostics. Historical configurations retain their original scoring conventions.

A Related Work

Inference-time alignment and search. NUDGING and BlendIn steer generation with auxiliary guidance; SafeDecoding, SafeInfer, InferAligner, and DeRa provide complementary approaches to inference-time alignment (Fei et al., 2025; Gan et al., 2026; Xu et al., 2024; Banerjee et al., 2025; Wang et al., 2024b; Liu et al., 2024b). Reward-guided decoding and search methods such as ARGS, DeAL, and TreeBoN evaluate candidate continuations during generation (Khanov et al., 2024; Huang et al., 2025; Qiu et al., 2025). CARE uses monitoring, rollback, and introspection to revise a generation path (Hu et al., 2026). MVSD studies the setting in which sampled suffixes serve as evidence and a short opening is committed under a fixed completion policy. Full-response best-of-N is an additional practical comparator: its scored response is also its delivered response.

Calibration, evaluation, and safety benchmarks. Split-conformal and multi-output calibration provide simultaneous uncertainty tools (Angelopoulos et al., 2024b; Angelopoulos & Bates, 2023; Dheur et al., 2025). MVSD applies these tools to absolute and comparative predictions of one invariant execution target, then propagates their joint constraints through a difference graph. HarmBench, SALAD-Bench, StrongREJECT, and JustEval supply complementary benchmark endpoints (Mazeika et al., 2024; Li et al., 2024a; Souly et al., 2024), with XSTest and OR-Bench providing additional views of safety and over-refusal (Röttger et al., 2024; Cui et al., 2024). Adversarial robustness draws on GCG and PAIR attack protocols (Zou et al., 2023; Chao et al., 2025), and JailbreakBench provides a standardized evaluation harness (Chao et al., 2024). Automatic evaluation quality is characterized in SORRY-Bench, SoSBench, and TrustJudge (Xie et al.; Jiang et al., 2026; Wang et al., 2026; Zheng et al., 2023; Wang et al., 2024a). Generated context can affect a fitted model through distraction, positional sensitivity, and shortcuts (Liu et al., 2024a); model-based policy optimization likewise studies how to use generated rollouts effectively.

Preference learning and inference-time feedback. Direct Preference Optimization provides a preference-based training objective (Rafailov et al., 2023), while RewardBench evaluates reward models used to rank language-model outputs (Lambert et al., 2025). These works contextualize the preference critic that supplies MVSD's reference ranking. RAIN explores self-evaluation and rewind during generation (Li et al., 2024b); MVSD instead uses sampled continuations to constrain the outcome of a committed opening.

Rubric-based and specialized evaluation. G-Eval studies structured evaluation with language models (Liu et al., 2023). Prometheus and JudgeLM develop trained evaluators, and PandaLM studies automatic comparison for instruction-tuning optimization (Kim et al., 2024; Zhu et al., 2025; Wang et al., 2024c). These approaches motivate explicit evaluation dimensions for helpfulness and task completion. Llama Guard and ShieldGemma provide dedicated input/output safety and content-moderation models (Inan et al., 2023; Zeng et al., 2024), complementing the general quality rubrics used in benign-request evaluation.

Evaluation bias and scope. Length-controlled AlpacaEval addresses response-length effects in automatic comparison (Dubois et al., 2024); this motivates interpreting depth and engagement alongside response budgets. Broader accounts of LLM-as-a-judge examine its opportunities and challenges (Li et al., 2025), and work on the limits of scalable evaluation studies the statistical value of model-based judgments (Dorner et al., 2025). Together, these perspectives motivate keeping evaluator scores, preference rankings, and execution-loss labels explicit in the evaluation protocol.

Risk control and selective prediction. Conformal Risk Control extends calibration to risk objectives, while Conformal Language Modeling applies conformal methods to language generation

(Angelopoulos et al., 2024a; Quach et al., 2024). Selective classification and SelectiveNet study the relationship between prediction coverage and risk (Geifman & El-Yaniv, 2017; 2019). MVSD uses a qualification rule with a reference-action fallback, so its coverage measures how often the controller can support a qualified action while still delivering an answer on every request.

Adaptive robustness. Simple adaptive jailbreak attacks provide a complementary perspective on robustness beyond fixed attack transformations (Andriushchenko et al., 2025). They motivate the feedback-attack analysis in Appendix C.5, alongside the fixed-attack benchmarks in the main evaluation.

B Data Composition, Independence, and Evaluation Coverage

The new study records source-family independence before fitting or calibration. Paraphrases, attacks, and resampled branches inherit their source family. Historical row counts below identify reference protocols; they are not confirmation-family counts. Benchmark populations and calibration partitions serve different purposes. The decision-study protocol specifies 1,000 calibration families and a 4,061-family precision target per confirmation population (Appendix I); these are not additional rows of each benchmark. The historical SALAD panel was previously exposed. Family grouping keeps related attacks together in paired comparisons.

Table A1: Benchmark populations and evaluation units.

Population	Evaluation coverage	Grouping unit
HB fixed attacks	960 answers	240 behaviors
SALAD attack-enhanced	5,000 answers	2,367 source questions
SR static five	1,565 answers	313 behaviors
JE quality / safety	800 / 200 requested answers	Request

B.1 Candidate opportunity and selection outcomes

Table A2: Candidate opportunity for binary harmful losses. Entries follow the loss definition and are mathematical identities, not estimated stratum frequencies.

Candidate bank	Oracle loss	Selected loss	Candidate-relative regret
All safe	0	0	0
Mixed	0	$q_{\hat{a}} \in \{0, 1\}$	$q_{\hat{a}}$
All unsafe	1	1	0

Only mixed banks permit a selector to change binary safety. All-unsafe banks retain loss one despite zero regret. The primary outcome comparison includes all banks; benign requests retain their separate graded-loss endpoint.

B.2 Scoring coverage and valid populations

Table A3: Aggregate Target scoring coverage across four generators on HB, SALAD, and SR.

Method / dataset	Requested	Generated	Valid scored	Unknown
Target, four generators	14,900	14,900	14,900	0

The aggregate covers 3,725 scored answers per generator; paired comparison populations are defined separately. Unresolved labels remain unknown. The JustEval configuration-specific coverage is given in Appendix D.

Table A4: Primary fixed-attack comparison. One S6 configuration is used throughout this table. ASR is lower-is-better; all rows have zero unknown scores.

Method	Unsafe answers (%)	Broken behaviors (%)
Target	3.23	10.42
BlendIn	2.92	9.58
NUDGING	6.88	21.25
InferAligner-simple (R)	2.19	8.33
SafeDecoding (A)	2.81	9.17
SafeInfer (A)	2.60	8.75
DeRa-SFT (A)	5.73	21.25
DeRa-DPO (A)	13.85	35.83
TreeBoN-DPO (A)	3.75	12.08
CFGD S6 adaptive	1.67	5.83

R: simple-variant reimplementation. A: component/expert adaptation. All methods use the same 240 behaviors and four fixed conditions; shared output budgets do not imply matched computation. Descriptive results after exploratory method development.

C Complete Benchmark Results and Historical References

These are exposed historical CFGD measurements, not results of the new evidence heads. The supplementary records retain the corresponding configurations and evaluation protocols.

C.1 Controller and evaluation units

The historical Qwen3.5-9B/Qwen3.5-4B system used $t = 0$, $K = B = 4$, $M = 2$, the five-horizon panel and 512-token greedy execution. It averaged two trace-wise preference complements, used the first preference gap ≥ 0.2, and otherwise chose at horizon 64. The dispersion threshold 1.0 is nonbinding because two scores in $[0, 1]$ have population standard deviation at most 0.5. Materializing the complete panel costs up to 512 sampled future tokens and ten critic calls, excluding proposals/completion; no early-stop saving follows.

Historical supervision used 200/40 train/validation anchors, 32 paired rollouts plus 32 for ambiguous cases, hard winners with positive one-sided 95% bootstrap bounds, and otherwise soft winner distributions with maximum mass at least 0.50 (zero weight below it). Permutations moved candidate labels together. Relative preference is not absolute risk. No original grid candidate passed its validation gate: mean regret was 0.02860 [0.00759, 0.06899], correct selection 0.325 [0.175, 0.475], and recovery 0.6065 [0.0601, 0.9023]. These stored intervals are not execution-safety guarantees.

HB uses 240 behaviors and four fixed attacks (960 answers). Response ASR averages attack indicators; behavior ASR counts any success per behavior. Transferred GCG is not defense-aware optimization. A separate direct panel has 60 validation plus 240 test behaviors. The older SALAD panel has 1,200 queries; the later 5,000-item panel is separate. SR averages harmfulness over 313 behaviors and five conditions. PAIR-style feedback attacks allow 25 queries per behavior with defense-dependent realized prompts (Chao et al., 2025). R denotes reimplementation, A adaptation, and missing values unavailable scores.

C.2 Fixed attacks and supplementary safety

Adaptive CFGD has 16 unsafe HB answers and 14 compromised behaviors versus Target's 31 and 25. Its per-condition changes are $(-3, -8, -2, -2)$ versus Target and $(-1, -8, +4, 0)$ versus InferAligner: lower aggregate ASR is not attack-wise dominance. On the older SALAD panel, safe counts are 1,097 (CFGD), 1,039 (Target), 1,050 (BlendIn) and 1,101 (InferAligner). On SR, CFGD's 0.02748 exceeds InferAligner's 0.01750 despite improving on the other three original comparators.

C.3 Paired SALAD statistics

Paired gains equal $(n_+ - n_-)/1200$; for Target, $(88-30)/1200 = 4.83$ percentage points. Pointwise bootstrap intervals and Holm tests retain their original seven-comparison family and exploratory

Table A5: Fixed-attack unsafe-answer rates (%); each condition covers 240 behaviors.

Method	AIM	Base64	GCG transfer	JSON injection
Target	1.25	4.58	4.58	2.50
BlendIn	0.00	3.33	4.17	4.17
NUDGING	5.42	3.75	11.25	7.08
InferAligner-simple (R)	0.42	4.58	2.08	1.67
SafeDecoding (A)	1.25	2.08	1.67	6.25
SafeInfer (A)	0.83	3.75	4.17	1.67
DeRa-SFT (A)	4.17	0.00	2.92	15.83
DeRa-DPO (A)	30.83	2.50	15.83	6.25
TreeBoN-DPO (A)	0.42	3.33	8.33	2.92
CFGD S6 adaptive	0.00	1.25	3.75	1.67

Table A6: Additional safety endpoints. Each SALAD row covers 1,200 queries; each available StrongREJECT row covers 313 behaviors under five fixed conditions.

Method	SALAD safe (%) ↑	StrongREJECT mean ↓
Target	86.58	0.03158
BlendIn	87.50	0.02951
NUDGING	87.25	0.03111
InferAligner-simple (R)	91.75	0.01750
SafeDecoding (A)	86.42	–
SafeInfer (A)	88.33	–
DeRa-SFT (A)	86.58	–
DeRa-DPO (A)	76.67	–
TreeBoN-DPO (A)	86.00	–
CFGD S6 adaptive	91.42†	0.02748

†: historical S6 artifact, retained separately from the expansion panel; this cell is not a result for all S6 horizon variants. StrongREJECT is a continuous score, not ASR. – indicates unavailable verified scores in the snapshot. InferAligner-simple StrongREJECT is rounded in its retained completion report.

Table A7: Retained S6 paired comparisons on 1,200 SALAD items. Intervals are pointwise paired-bootstrap intervals (100,000 resamples); Holm adjustment applies to the original seven comparisons only. Exploratory selection bias remains.

Comparator	Gain (pp)	95% CI (pp)	Improved/worsened (%)	Holm p
S0	9.17	[7.08, 11.25]	11.67/2.50	2.41×10^{-17}
S1	3.08	[1.50, 4.67]	5.58/2.50	3.06×10^{-4}
S2	3.25	[1.67, 4.92]	5.92/2.67	3.06×10^{-4}
S3	6.58	[4.67, 8.50]	9.17/2.58	8.63×10^{-11}
Target	4.83	[3.08, 6.58]	7.33/2.50	4.28×10^{-7}
BlendIn	3.92	[2.25, 5.67]	6.67/2.75	3.44×10^{-5}
NUDGING	4.17	[2.50, 5.83]	6.67/2.50	8.07×10^{-6}

selection bias. They do not establish superiority over the subsequently added InferAligner-simple. Additional views or transformations are not independent observations.

C.4 Utility and horizon sensitivity

Joint safety/helpfulness is the within-answer indicator of safety and helpfulness ≥ 2. XSTest pools 250 safe/200 unsafe requests; JE pools 800 quality/200 safety requests; OR pools 990 Hard/492 Toxic requests. These pooled scores are not safe-only refusal rates or JE five-dimensional quality. CFGD's joint XSTest/JE scores are below Target; OR content safety is 100%, but 1,465/1,482 joint successes are one below Target. Human agreement and utility noninferiority are unestablished. Adaptive and H8 tie on fixed HB; H8 has lower SR loss, while H4 is best among these CFGD variants on feedback attacks. No horizon dominates all endpoints, and these aggregates do not isolate supervision, augmentation or adaptive selection.

Table A8: Supplementary joint safety-and-helpfulness diagnostic (local 70B judge). Higher is better under this rubric.

Method	XSTest joint	(%)	JustEval joint	(%)
Target	435/450	96.67	987/1000	98.70
BlendIn	435/450	96.67	989/1000	98.90
NUDGING	432/450	96.00	983/1000	98.30
InferAligner-simple (R)	436/450	96.89	725/1000	72.50
SafeDecoding (A)	442/450	98.22	987/1000	98.70
SafeInfer (A)	437/450	97.11	993/1000	99.30
DeRa-SFT (A)	367/450	81.56	850/1000	85.00
DeRa-DPO (A)	440/450	97.78	986/1000	98.60
TreeBoN-DPO (A)	439/450	97.56	987/1000	98.70
CFGD S6 adaptive	430/450	95.56	985/1000	98.50

All 450 XSTest items and all 1,000 JustEval items are included. These mixed-subset scores are not safe-only answer rates or canonical five-aspect JustEval results. Listed rows have zero unknown joint scores. The retained local-judge identity is Meta-Llama-3-70B-Instruct.

Table A9: Horizon sensitivity across protocols. Every numeric column is lower-is-better. Adaptive refers to the defense controller, not the attacker.

S6 variant	Fixed resp. ASR	Fixed beh. ASR	SR mean	Adaptive ASR
Fixed $H = 4$	2.08	6.67	0.02935	18.33
Fixed $H = 8$	1.67	5.83	0.02559	21.25
Fixed $H = 16$	1.88	6.67	0.02835	22.92
Fixed $H = 32$	1.88	6.67	0.02582	18.75
Fixed $H = 64$	1.77	6.25	0.02583	21.67
Adaptive	1.67	5.83	.02748	20.42

ASRs are percentages. Fixed attacks: 960 answers / 240 behaviors. SR: StrongREJECT static5, 1,565 answers. Adaptive attacker: 240 behaviors, maximum 25 queries. These are exploratory variants, not interchangeable main-result cells.

C.5 Feedback attack analysis

The full budget table includes the maximum-budget comparison: adaptive CFGD reaches 49/240 compromises versus Target 54, NUDGING 72, DeRa-SFT 19, SafeDecoding 21, InferAligner 29 and same-critic BoN 39. H4 reaches 44. Shared query limits do not imply equal realized attacks or compute.

Reproduction and preparation. The evaluation summaries retain method configurations, scoring coverage, and negative results. An AI assistant was used for manuscript editing, algebra checks, and figure preparation.

D Historical Best Variants and Supplementary Benchmarks

The baseline measurements appear once in Table 1. Best-observed CFGD choices are adaptive/H8 on HB (1.67% response and 5.83% behavior ASR), H16 on SALAD (92.44%), and H8 on SR (0.02559). JE uses the two fixed configurations below. These exploratory choices do not form one deployable configuration; the cyclic refinement is separate.

D.1 H16 cyclic-averaging historical regression

The retained September 25 report keeps H16, K4/B4/M2, and native512 completion, averaging critic scores over four cyclic candidate orders. It records 4,646/5,000 safe answers (92.92%) versus 4,622 (92.44%) for original H16: 49 improvements, 25 regressions, and net +24 (+0.48 percentage points). The source-family bootstrap interval is [+0.14, +0.825] points over 2,367 source questions. This is exposed historical regression, unadjusted for prior search and follow-up decisions; it does not

Table A10: All retained adaptive-attack methods and variants. Budget columns give attack-success rates (%) over 240 behaviors; the final column repeats ASR at budget 25.

Method	1	5	10	15	20	25	ASR
SafeInfer (A)	3.33	7.08	13.33	15.83	17.92	18.75	18.75
NUDGING	5.00	13.75	21.25	27.50	29.58	30.00	30.00
DeRa-DPO (A)	6.67	22.92	34.17	37.92	41.67	42.08	42.08
InferAligner-simple (R)	2.50	6.25	8.75	11.25	12.08	12.08	12.08
CFGD S6 adaptive	0.83	8.33	13.33	16.67	19.58	20.42	20.42
S6 fixed H4	1.25	5.42	10.42	14.17	17.50	18.33	18.33
S6 fixed H8	0.42	7.08	11.67	15.42	20.42	21.25	21.25
S6 fixed H16	0.83	7.50	12.08	16.67	21.67	22.92	22.92
S6 fixed H32	0.83	7.08	12.92	15.42	18.33	18.75	18.75
S6 fixed H64	0.83	9.17	14.58	18.75	21.25	21.67	21.67
S7 anchor	0.83	8.33	12.92	16.25	19.17	20.00	20.00
S7 tau=0.025	0.83	8.33	12.92	16.25	19.17	20.00	20.00
S7 tau=0.05	0.83	8.33	12.92	15.83	18.75	19.58	19.58
S7 tau=0.1	1.25	8.75	13.33	16.25	19.17	20.00	20.00
BoN (same critic)	2.92	6.25	10.42	13.75	14.58	16.25	16.25
TreeBoN-style (same critic)	2.08	7.08	11.25	16.25	19.17	19.58	19.58
TreeBoN-DPO (A)	3.33	12.92	20.42	23.75	25.42	26.25	26.25
DeRa-SFT (A)	1.25	3.75	5.00	7.08	7.50	7.92	7.92
Target	3.33	10.83	15.83	20.83	22.08	22.50	22.50
BlendIn	2.50	10.42	15.83	20.42	21.67	23.33	23.33
SafeDecoding (A)	1.25	5.00	6.67	7.50	8.33	8.75	8.75

PAIR-style feedback attacks with official HB scoring; 240 behaviors per method. Shared maximum query budgets do not imply identical realized prompts or query totals.

establish independent confirmation or CFGD certification. Critic calls increase from two to eight; full latency and benign utility remain unmeasured for this change.

D.2 All retained fixed variants on the safety panels

Table A11: Complete 21-method HB/SALAD historical panel and available SR entries. Best-variant selection is exploratory; protocols and denominators differ across columns.

Method	HB ASR (%)	SALAD safe (%)	SR mean
S6-H16	1.88	92.44	0.02835
S6-H32	1.88	91.96	0.02582
S7 tau=0.025	1.77	91.44	—
S7 anchor	1.77	91.44	—
InferAligner-simple (R)	2.19	91.42	0.01750
S7 tau=0.05	1.77	91.40	—
S6-adaptive	1.67	91.36	0.02748
S6-H8	1.67	91.30	0.02559
S7 tau=0.1	1.77	91.26	—
S6-H64	1.77	91.04	0.02583
TreeBoN-style (same critic)	2.08	89.50	—
S6-H4	2.08	89.02	0.02935
BoN (same critic)	1.46	88.26	—
SafeInfer (A)	2.60	87.68	—
blendin	2.92	87.40	0.02951
TreeBoN-DPO (A)	3.75	86.86	—
target	3.23	86.76	0.03158
nudging	6.88	86.76–86.78	0.03111
SafeDecoding (A)	2.81	86.08–86.10	—
DeRa-SFT (A)	5.73	85.78	—
DeRa-DPO (A)	13.85	75.92	—

HB has 240 behaviors and four fixed attacks. SALAD has 5,000 identical IDs per method; the NUDGING/SafeDecoding ranges each reflect one unknown label. SR uses 313 behaviors and five conditions; — is an unavailable verified value. H16 cyclic averaging is a later regression, reported separately.

D.3 Fixed optimized JE configurations and scoring coverage

Table A12: Two fixed JE configurations from September 17. These are separate from legacy H16 and from CFGD.

Configuration	Help.	Fact.	Depth	Clarity	Engage.	Safety
HFU-v16-P050	4.9098	4.8882	4.3787	4.9822	4.0343	5.0000
S6-P050	4.9034	4.8780	4.3863	4.9771	4.0610	5.0000

Both optimized rows use 787/800 valid quality scores and 200/200 safety scores, with a 2048-token budget. Historical September 11 quality counts are Target 787, BlendIn 787, NUDGING 781, InferAligner 788, SafeDecoding 783, SafeInfer 780, DeRa-SFT 776, DeRa-DPO 782, TreeBoN-DPO 777, BoN 789, and TreeBoN-style 788. Safety counts are 200 except TreeBoN-style (199). Mean safety ratings are 4.9950 for NUDGING, 4.9750 for DeRa-DPO, and 5.0000 for the other listed baseline/control and optimized rows; these are ratings, not safety percentages. The original answers and unresolved labels are preserved; budget and valid-sample differences remain explicit.

D.4 OR and XSTest supplementary observations

In the retained September 11 local-judge snapshot, S6 adaptive reaches 100% OR content safety; the same-critic TreeBoN-style arm has 99.39% joint safety/helpfulness versus SafeInfer's 99.19%. These are different endpoints and methods. On XSTest, S6-H32 content safety is 99.78% (tied with same-critic BoN), versus SafeDecoding's 99.56%; the candidate pool's best joint score is BoN's 96.67%, below SafeDecoding's 98.22%. Complete refusal/answering categories are retained in the evidence archive. These supplements do not change the four main benchmarks or establish uniformly improved benign utility.

E Evidence Contributions and the Decision Ledger

For each population separately, let $\Delta_n = q_{a_0,n} - q_{a_C,n}$. Define positive and negative contributions as $G_+ = N^{-1}\sum_n \max(\Delta_n, 0)$ and $G_- = N^{-1}\sum_n \max(-\Delta_n, 0)$. The all-anchor identity is $\bar{\Delta} = G_+ - G_-$. Counts of changed actions alone do not determine these contributions. These differences reproduce Figure 4. They do not identify the separate G_+ and G_- terms, which require paired outcomes. Harmful and benign endpoints are not pooled.

Table A13: Endpoint differences computed from the reported means on the common bank.

Endpoint	H16	Full composition	H16 minus full
Harmful regret	0.00554	0.00485	+0.00069
Benign loss	0.00523	0.00514	+0.00009

E.1 Multi-view support and constraint deletion

Enumerate all 31 nonempty horizon subsets in the full five-view diagnostic. Record separately the minimal views needed to qualify the same action and to support a positive improvement bound for the fixed action pair. Then rerun the selector after deletion and record changed actions, improvements, regressions, and ties. Keep horizon identities explicit $(4, 8, 16, 32, 64)$; these identify evidence views, not two algorithmic stages. In the optional stopping extension, enumerate reached views only; eligible checks begin at 16. An empty original panel remains a conflict, rather than being repaired by post hoc deletion. Full details are in Appendix I.

F Calibration and Component Results

The calibration summary separates simultaneous coverage, qualification, conditional violations, and interval width. Conditional qualified rates and the marginal guarantee use different denominators.

Table A14: Reported head-configuration diagnostics. Rates are percentages; interval width is on the executed-loss scale.

Variant	Coverage (%)	Qualification (%)	Qualified violation (%)	Width
Absolute-only	89.6	47.8	9.2	0.087
Derived differences	90.3	68.2	8.9	0.071
Direct joint heads	90.2	100.0	9.8	0.054

The component protocol distinguishes unit from development scales, coordinatewise from coherent bounds, and current from cumulative evidence. Attribution uses fixed intervals, while practical variants receive independent calibration. Direct and derived heads share the information and labels defined in Appendix M; their interval widths and executed outcomes are separate endpoints.

G Configuration Stability and Computational Profiles

Table A15: Reported latency and outcomes used in Figure 6. Harmful and benign request populations are shown separately.

Population	Configuration	Latency (ms)	Regret / loss
Harmful	H16	1,247	0.00554
Harmful	Greedy-prefix evidence	1,384	0.00512
Harmful	Full MVSD	1,458	0.00485
Harmful	Stopping extension	1,203	0.00501
Benign	H16	1,231	0.00523
Benign	Full MVSD	1,442	0.00514

The full panel uses ten logical critic comparisons versus two for H16. Generation, batching, and final completion also contribute to latency, so this call ratio does not imply a fivefold end-to-end slowdown. The plotted values are the reported means; no uncertainty intervals or additional timing runs are inferred from them. Separate hardware, batching, and token profiles are needed for transfer to a different deployment setting.

Supporting robustness protocols retain seed, horizon, candidate-order, and suffix-visibility checks. Feedback attacks and OR/XSTest measurements appear in Appendices C and D; the visibility studies are specified in Appendix Q.

H Decision-Study Design

This section specifies the common-bank comparison and the supporting diagnostics. Source-family splits, candidates, executed answers, labels, heads, budgets and contrasts define each comparison. The primary bank uses $K = B = 4$, $M = 2$, $t = 0$, and $L = 64$ on the five-horizon grid. Evaluate the entire population, including all-safe/all-unsafe/mixed candidate sets, conflicts and fallback; benign graded losses form a separate endpoint.

The controls in Table A16 connect executed gain to retained constraints, simpler selectors, learned integration, and computation. Appendix I defines their comparison protocol; Appendix Q defines the supporting branch diagnostic.

I Action-Correction Evaluation Protocol

Population and contrasts. The primary contrasts compare H16, direct prefix veto, representation fusion, and a development-selected fixed horizon against full MVSD. Harmful-regret and benign-loss contrasts form separate multiplicity families; a combined claim uses joint correction over all eight. Long-only minimax is an additional diagnostic. Paired means use independent source families, with related traces and attack transforms grouped together. Corpus versions, splits, and eligibility are fixed before evaluation.

Table A16: Each control challenges a link in the same decision argument. Attribution shares intervals; practical comparisons independently train/calibrate frozen variants.

Control	Information or operation	Question for action correction
Short / selected horizon	One development-frozen view	Does fixed truncation suffice?
Legacy H16	Original critic and preference order	Does the evidence module improve its backbone?
Long-only	Latest prediction	Does the final view suffice?
Direct / propagated veto	Prefix exclusions, then H16 ranking	Does simple preservation suffice?
Learned fusion	All horizon scores or representations	Can learned integration recover the gain?
Absolute / derived heads	Losses, with or without derived differences	Is a learned comparison head useful?
Greedy-prefix / full greedy	Execution-consistent evidence	Is sampled or partial evidence worthwhile?
MVSD	Cumulative joint constraints	Does composition yield net decision value?

Selectors and matched information. MVSD uses Section 2. Long-only minimizes $\max_j g_{ij,L}$ with fixed action-key ties. Choose the fixed horizon by minimum family-averaged development loss, breaking ties toward the shorter horizon. Direct and propagated veto use Appendix S, reference ranking and the same a_0 fallback. Fit horizon-specific critics as a control for unequal horizon fit. Attribution shares a fitted model's full-path intervals; deployment arms independently train/calibrate each declared model. Unconstrained rankings have no qualification guarantee.

Fusion architectures. Score fusion concatenates each candidate's five absolute scores and pairwise rows, with horizon/candidate masks, into $5(K+1)$ inputs. A two-layer width-256 GELU network feeds sigmoid absolute and antisymmetric comparison heads, trained on the same q, d loss with base scores frozen. The primary representation-fusion control projects each original trace-wise $u_{i,h}^{(m)}$ to width 256, adds a horizon embedding, and learns candidate-specific softmax attention over all ten horizon–trace records. Apply the same prediction heads to the pooled vector. Train pooling/projection/heads, keeping the common encoder frozen; label end-to-end fusion separately.

Fusion shares source families, seeds, optimizer, and development selection with Appendix M. Training comparisons account for parameters, updates, tokens, and FLOPs: fusion uses one pooled record per anchor, whereas MVSD averages horizon losses. A stopping variant masks unreached views and calibrates its full eligible path.

Two removal audits. The first removal audit fixes (a_0, a_C) and calibrated endpoints, then enumerates the 31 nonempty horizon subsets to identify qualification and positive-Γ support. The second reruns selection after deletion to measure action and loss changes. Absolute and pairwise constraints are removed separately; all subsets retain the unit cube, and original conflicts remain conflicts. Stopping diagnostics use reached views only. Multiple minimal subsets permit shared rather than unique attribution.

Ledger and practical controls. The decision ledger links each source family to executed losses, predictions, calibrated bounds, selected actions, qualification, Δ, Γ, and cost. Positive and negative contributions obey Appendix E, separately for harmful and benign populations. The supporting witness diagnostic uses $h = 8$ and $\delta = 0.10 > \varepsilon = 0.05$.

Execution-consistent controls compare sampled and greedy-prefix evidence at matched visible lengths and generation budgets. Full-candidate scoring selects among completed answers using a deployment-available predictor and delivers the selected answer directly. Latency includes valid prefix reuse, fallback, and failures; the all-action loss oracle is used only for offline evaluation.

Precision and interpretation. For m family-averaged contrasts in $[-1, 1]$, Hoeffding plus Bonferroni gives

$$N \geq \left\lceil 2\log(2m/\eta)/w^2 \right\rceil . \tag{11}$$

For $\eta = w = 0.05$, Equation 11 gives 4,061 independent families for $m = 4$ and 4,615 for $m = 8$, per evaluated population. Calibration uses a separate 1,000-family partition. These are precision requirements; additional transforms of one source family do not increase the independent sample size.

Executed gains, benign utility, and selector comparisons evaluate practical benefit. Fixed-interval removal links that benefit to composition; qualification and branch visibility are distinct diagnostic endpoints.

J Additional Learning and Controller Details

Shared information and head parameterization. Reuse the original critic input for each trace at horizon h. Its terminal input-token state $r_h^{(m)}$ follows all visible candidate blocks; no new delimiter or other trace is appended. Define $u_{i,h}^{(m)} = [r_h^{(m)}; e_i^{(m)}]$, where the one-hot $e_i^{(m)}$ is the candidate's slot under that call's original visit-0 permutation. Both heads receive this same contextual state and slot mapping. With a sigmoid absolute head and shared comparison MLP f_ϕ, set

$$g_{ij,h}^{(m)} = \tfrac{1}{2}\{\tanh f_\phi(u_{i,h}^{(m)}, u_{j,h}^{(m)}) - \tanh f_\phi(u_{j,h}^{(m)}, u_{i,h}^{(m)})\}. \tag{12}$$

The comparison input is $[u_i, u_j, u_i - u_j]$. Both heads use width-$w = 256$ tanh MLPs; for readout dimension D, their total parameter count is $4(D+K)w + 4w + 2$ including biases. Map predictions to original action identities before the mean in Equation 2. Averaging preserves antisymmetry, zero diagonal and the output ranges. It is also invariant to exchanging the two complete trace records; this does not imply invariance to reserializing candidates inside a critic input.

Equation 3 supervises the aggregated outputs, with unit weight per adjudicated anchor. The original backbone, adapter and preference head stay frozen. The same two heads are reused across all horizons and traces. Direct and derived controls share these readouts, mapping and aggregation; for the derived arm, $g_{ij,h} = v_{i,h} - v_{j,h}$, equivalently the mean of trace-wise differences. Graded benign losses distinguish regret effects from acceptability; with binary harmful loss a zero-loss action already has zero regret.

Population optimum and finite models. For unrestricted square-loss predictors with the same information F_h, the population optima satisfy

$$v_{i,h}^* = \mathbb{E}[q_i \mid F_h], \qquad g_{ij,h}^* = \mathbb{E}[q_i - q_j \mid F_h] = v_{i,h}^* - v_{j,h}^*. \tag{13}$$

The same identity holds conditional on a shared frozen representation. Trace-wise averaging is a restricted fitted class, and calibration is applied to its resulting predictions. The identity need not hold across different information sets or finite fitted classes. Direct comparisons have no inherent information advantage. Our strong derived control is trained with *both* terms of Equation 3, substituting $v_i - v_j$ for g_{ij} and backpropagating the pair loss through both scalar outputs. An absolute-loss-only derived arm isolates the effect of this training signal. Appendix L.6 proves the common optimum and states the objectives.

The finite-model hypothesis concerns residual tails useful for calibrated decisions. Propagation must offset joint-radius inflation (Proposition 3); validity alone establishes no learned-head advantage.

A direct learning-to-decision link. Let $\tilde{i}_h = \arg\min_i \max_j g_{ij,h}$, with fixed token-key tie breaking. If $e_h = \max_{i,j} |g_{ij,h} - d_{ij}|$, then

$$q_{\tilde{i}_h} - \min_j q_j \le 2e_h, \qquad \mathbb{E}R_E(\tilde{i}_h) \le 2\sqrt{P\,\mathbb{E}\mathcal{L}_{\text{pair,h}}}. \tag{14}$$

The expectation bound uses population pairwise error for the unconstrained minimax selector. Proposition 1 supplies the separate guarantee for the qualified controller. Both derivations appear in Appendix L.

Calibrating unequal error scales. Freeze the fitted model, positive scales σ_h^v, σ_h^g, and all design choices before opening an independent calibration set of n anchors. Scales are fitted on development

data only; Appendix M specifies a reproducible RMS rule. For each calibration anchor, materialize the nested evidence bank over $\mathcal{H} = \{4, 8, 16, 32, 64\}$ and compute

$$S = \max_{h \in \mathcal{H}} \max \left\{ \max_i \frac{|v_{i,h} - q_i|}{\sigma_h^v}, \max_{i<j} \frac{|g_{ij,h} - (q_i - q_j)|}{\sigma_h^g} \right\}. \tag{15}$$

Let c_α be the kth smallest calibration score, $k = \lceil (n+1)(1-\alpha) \rceil$, with $c_\alpha = +\infty$ when $k > n$. Set $r_h^v = c_\alpha \sigma_h^v$ and $r_h^g = c_\alpha \sigma_h^g$. This uses established split-conformal simultaneous coverage (Angelopoulos et al., 2024b; Dheur et al., 2025). Scaling allows different raw widths across heads and horizons; it does not guarantee narrower intervals than unscaled calibration. Its quality affects usefulness, not rank validity.

Exact bounds by difference constraints. The feasible set has only constraints of the form $z_i - z_j \le b$. Add node 0 with $z_0 = 0$; an edge $j \to i$ with weight b represents this inequality. Absolute intervals give edges to and from node 0. Shortest-path closure either detects infeasibility through a negative cycle or gives distances D_{ji} with

$$V_{i,h} = D_{0i}, \qquad U_{i,h} = \max_{1 \le j \le K} D_{ji}. \tag{16}$$

These are exact extrema for the stated feasible set. Standard all-pairs closure costs $O((K+1)^3)$ per reached horizon, with five nodes for $K = 4$. The contribution is using the invariant executed target to accumulate and propagate learned constraints for commitment; shortest-path algorithms and conformal rank coverage are established tools. Appendix L gives the graph construction, proof and numerical safeguards.

Qualification and gain. Section 4 separates unchanged qualification, certified correction, and changes with unresolved sign. Shared fallback gives $\Delta = 0$ on nonqualification. On simultaneous coverage, qualified regression is at most $\min(\beta, \varepsilon)$.

Core mechanism: complementary constraints across horizons. Table A17 constructs two graded-loss panels at $h_1 = 4$ and $h_2 = 8$, containing the same $q = (0.04, 0.40)$. Neither horizon alone qualifies action 1 at $\beta = 0.10$, $\varepsilon = 0.05$. The view at h_1 supplies the absolute anchor $z_2 \le 0.45$; the view at h_2 supplies $z_1 - z_2 \le -0.36$. Their composition gives

$$z_1 \le z_2 - 0.36 \le 0.09, \qquad U_1 = 0.$$

Removing either horizon, keeping only absolute intervals, or using direct comparisons at h_1 destroys this certificate. The example isolates cross-head and cross-horizon certification; the action-changing construction and common-bank comparison address the selected outcome.

Minimal subsets and first-stop logic. For a qualified action i, $m_i(h)$ is the smallest reached-view subset qualifying the same action under fixed intervals. Enumerating at most 31 subsets distinguishes current-view, prior-single-view, and genuine multi-view support. At a first eligible stop, an earlier eligible single view cannot be solely necessary: its earlier cumulative panel would already have triggered selection. Views 4 and 8 are ineligible stops, so either can suffice at the initial check 16. Qualification support, positive-Γ support, and actual changes after deletion remain separate diagnostics.

K Tabulated Mathematical Constructions

The following mathematical examples instantiate the mechanism in Section 4.

L Execution-Regret Proofs and Guarantee Boundaries

The statements here establish the guarantees of MultiView Safe Decodingunder the specified execution and calibration assumptions. Split-conformal rank coverage and simultaneous output regions are established tools (Angelopoulos et al., 2024b; Dheur et al., 2025); the task-specific construction is the joint regret/loss envelope used to select and stop on an executed action.

Evidence	Interval for z_1	Interval for z_2	$z_1 - z_2$	Closed V_1
View $h_1 = 4$	$[0, 0.30]$	$[0.15, 0.45]$	$[-0.60, 0.02]$	0.30
View $h_2 = 8$	$[0, 0.80]$	$[0.20, 1]$	$[-0.40, -0.36]$	0.64
Both	Intersect the two panels and propagate comparisons			**0.09**

Table A17: Mathematical example of complementary constraints. The certificate requires both horizons: the relation at h_2 combines with the absolute anchor at h_1 to establish acceptability. Widths are representable under the declared calibration form (Appendix L.10).

View	Absolute centers v	r^v	Pair centers (g_{12}, g_{13}, g_{23})	r^g
View $h_1 = 4$	$(0.30, 0.30, 0.05)$	0.30	$(0.20, 0.20, 0.48)$	0.50
View $h_2 = 8$	$(0.50, 0.50, 0.50)$	0.50	$(0.10, -0.76, -0.46)$	0.50

Table A18: Constructed action-gain example. Apply the stated radii symmetrically and intersect with $[0, 1]^3$; reverse pair centers by antisymmetry. The two-horizon rule gains 0.26 over reference action 2 and direct prefix veto. Setting $c_\alpha = 1$ and scales equal to these radii realizes the panels under the calibration parameterization.

L.1 Proof of the surrogate-to-regret bound

Suppress the horizon and let $i^* \in \arg\min_i q_i$. Since $d_{ij} = q_i - q_j$ and $d_{ii} = g_{ii} = 0$, $\max_j d_{ij} = q_i - \min_j q_j$. If $e = \max_{ij} |g_{ij} - d_{ij}|$, then $\max_j g_{i^* j} \le e$. Minimax selection implies $\max_j g_{\tilde{i}j} \le e$, and hence

$$R_E(\tilde{i}) = \max_j d_{\tilde{i}j} \le \max_j g_{\tilde{i}j} + e \le 2e.$$

Antisymmetry gives $e^2 \le \sum_{i<j} (g_{ij} - d_{ij})^2 = P\mathcal{L}_{\text{pair}}$. Jensen's inequality yields Equation 14. The bound is clipped by the trivial maximum regret one if desired. It does not assume pairwise predictions are transitive. A common additive shift in all true action losses leaves d_{ij} unchanged; this is why a separate absolute-loss head is necessary. Low empirical training loss does not establish the population error in the bound.

L.2 Proof of Proposition 1

Condition on the fitted heads, positive development scales, horizon set, rubric, proposal/evidence policies, and tie-breaking rules. Calibration and new-anchor scores in Equation 15 are exchangeable. Conservative rank coverage gives $\Pr(S_{n+1} \le c_\alpha) \ge 1 - \alpha$. On this event, every absolute error is bounded by r_h^v and every difference error by r_h^g, at every horizon. Hence the same true vector q belongs to all $\mathcal{Q}_h$ and their intersections $\mathcal{Q}_{\le h}$. These sets are nonempty, and maximizing over them gives $q_i \le V_{i,h}$ and $q_i - \min_j q_j \le U_{i,h}$ simultaneously. The data-dependent qualified action and horizon therefore satisfy both tolerances on this event. This proves Equation 8. Nested-prefix dependence is allowed: horizons are coordinates of one anchor score, not independent calibration samples.

When $c_\alpha = +\infty$, the cube $[0, 1]^K$ supplies trivial bounds of one for $K \ge 2$, and the strict tolerances preclude qualification. When an intersection is empty, it cannot contain the true vector and thus occurs outside the simultaneous coverage event. The controller rejects certification rather than maximizing over an empty set. This rule neither repairs a coverage failure nor makes the fallback safe.

The guarantee averages over calibration and a new anchor, conditional on the fitted assets. It does not give a high-probability conditional risk bound for each realized calibration bank. If $p_C = \Pr(C) > 0$, $\Pr(\text{violation} \mid C) \le \min(1, \alpha / p_C)$, not α. Calibration reuse for variant selection is outside the argument; confirmation must independently estimate coverage, violations, conflicts and fallback outcomes.

L.3 Proof of Proposition 2 and its limits

For $h' \ge h$, $\mathcal{Q}_{\le h'} \subseteq \mathcal{Q}_{\le h}$. If both sets are nonempty, maximizing either z_i or $z_i - \min_j z_j$ over the smaller set cannot increase its value. The same argument applies when comparing full constraints

with a relaxation obtained by dropping pairwise constraints or earlier horizons. Thus, on a nonempty path, an action that meets both tolerances continues to meet them as additional evidence is revealed. The primary controller selects at L. In the optional extension, the first qualifying eligible check is the earliest certificate under that check schedule, not an optimal stopping theorem over alternative computation policies.

For a current horizon, direct coordinatewise bounds from the same intervals are

$$\bar{V}_{i,h} = \min(1, v_{i,h} + r_h^v), \qquad \bar{U}_{i,h} = \min\{1, \max(0, \max_{j \neq i}(g_{ij,h} + r_h^g))\}.$$

Every feasible vector obeys these inequalities; therefore the cumulative bounds satisfy $V_{i,h} \leq \bar{V}_{i,h}$ and $U_{i,h} \leq \bar{U}_{i,h}$. The comparison holds for the same calibrated intervals on a nonempty intersection. It does not compare calibration quantiles: a joint-head score can have a larger quantile than an absolute-only score. Independently trained or recalibrated controls may outperform the joint method. Development scaling likewise need not improve raw widths. Those are empirical questions.

L.4 Graph construction and exactness

For finite c_α, form the cumulative endpoints

$$l_i = \max\{0, \max_{t \leq h}(v_{i,t} - r_t^v)\}, \qquad u_i = \min\{1, \min_{t \leq h}(v_{i,t} + r_t^v)\},$$
$$b_{ij} = \min_{t \leq h}(g_{ij,t} + r_t^g), \qquad i \neq j,$$

where t ranges over reached horizons. Antisymmetry supplies both directions of each pair interval. Initialize edges $0 \to i$ with weight u_i, $i \to 0$ with weight $-l_i$, and $j \to i$ with weight b_{ij}; initialize diagonal distances to zero and other missing edges to infinity. The cube bounds make every node reachable from every other node. Use standard all-pairs shortest-path closure. A negative cycle implies infeasibility, since summing its inequalities gives $0 < 0$. Conversely, with no negative cycle, distances from node 0 give a feasible potential, so infeasibility is detected exactly.

Every path from j to i sums valid inequalities and yields $z_i - z_j \leq D_{ji}$. The shortest-path bound is attainable: for fixed j, set $z_k = D_{jk} - D_{j0}$. Triangle inequalities ensure every edge constraint, $z_0 = 0$, and the cube bounds. Since $D_{jj} = 0$, this potential attains $z_i - z_j = D_{ji}$. Hence

$$\max_{z \in \mathcal{Q}_{\leq h}} z_i = D_{0i}, \qquad \max_{z \in \mathcal{Q}_{\leq h}} \max_j (z_i - z_j) = \max_j D_{ji}.$$

The maximizing vector may differ for each i, j; the finite maximum can nevertheless be interchanged with maximization over the common compact feasible set. Including $j = i$ ensures a nonnegative regret bound. This proves Equation 16 without assuming transitive raw predictions.

The theorem uses exact arithmetic. A reference implementation must retain the represented predictions and scales, round interval endpoints outward, and use exact rational closure or certified arithmetic. Ordinary floating-point rounding must not turn an uncertain bound into a certificate. If numerical feasibility or a threshold comparison cannot be certified, mark the anchor uncertified and include its cost and outcome. Once a cumulative set is empty, do not drop inconvenient constraints to restore qualification. The runner retains the conflicting constraints and returns the original H16 action without qualification.

L.5 Net gain after separate calibration

Use the graph before merging parallel edges from different horizons. An absolute upper edge $0 \to i$ has center $v_{i,t}$ and scale σ_t^v; a lower edge $i \to 0$ has center $-v_{i,t}$ and the same scale. A comparison edge $j \to i$ has center $g_{ij,t}$ and scale σ_t^g. Cube edges have centers one ($0 \to i$) or zero ($i \to 0$), with scale zero. Summing centers and scales on a path gives $\mu_p + c\lambda_p$. All edges use reached horizons only. Parallel-edge minimization and shortest-path closure recover the original controller exactly.

The absolute-only score omits pairwise residuals but retains all absolute coordinates and horizons. Each score receives its own rank quantile at the same α. Both calibrations use banks independent of training/development and test; the banks may be shared across these frozen variants. If the same anchors and rank are used, $c_J \geq c_A$ follows from coordinate inclusion. With different banks no such ordering is required. Proposition 3 uses $\Delta = (c_J - c_A)_+$ to cover both cases. An infinite quantile is excluded from the proposition and handled by the original fallback rule.

Proof. At the joint radius, the absolute path costs at most

$$\mu_p + c_J\lambda_p \le \mu_p + c_A\lambda_p + \Delta\lambda_p \le \beta.$$

Closure gives $V_i^J \le \beta$ and $U_i^J \le \varepsilon$ from the competitor paths. Since $A_i > \beta$, the absolute-only model cannot certify that action. This proves the stated action-level gain. It implies additional qualification by this horizon only if the absolute-only controller has no qualifying action at any reached horizon. Evaluating a comparison path at c_A is an algebraic reference cost, not a claim that its pairwise intervals are calibrated at c_A. The conclusion needs the actual joint quantile and nonempty joint set. It does not impose a new probabilistic guarantee beyond Proposition 1 or predict how often the condition holds.

A strictly positive calibration cost. Take one horizon with $v = (0.15, 0.30)$, $\sigma^v = 0.15$, $g_{12} = -0.38$, $\sigma^g = 0.02$, and illustrative quantiles $c_A = 0.8, c_J = 1$. The absolute-only box is $[0.03, 0.27] \times [0.18, 0.42]$; neither action passes $\beta = 0.10$. The joint box before comparisons is $[0, 0.30] \times [0.15, 0.45]$, with $q_1 - q_2 \in [-0.40, -0.36]$. Path $0 \to 2 \to 1$ has $\mu_p = -0.08$, $\lambda_p = 0.17$, and cost 0.09 at c_J. Its margin at c_A is 0.044, exceeding inflation 0.034. Regret for action one is zero throughout the joint set. Both sets contain $q = (0.04, 0.40)$, so the gain does not rely on a coverage failure of either comparator. These are constructed feasible intervals and stipulated quantiles, not calibration measurements. If the inflation exceeds the available path margin, this witness no longer certifies the action; another path may still do so.

L.6 Equal-information square-loss boundary

Fix a horizon and let $\mathcal{I}$ be the information common to both predictors, including candidate identities. It may be $\sigma(F_h)$ or the information in a frozen common readout. Let $\mu_i = \mathbb{E}[q_i \mid \mathcal{I}]$. Conditional square-loss decomposition gives

$$\mathbb{E}[(v_i - q_i)^2 \mid \mathcal{I}] = (v_i - \mu_i)^2 + \mathrm{Var}(q_i \mid \mathcal{I}),$$

with the analogous identity for $q_i - q_j$, whose conditional mean is $\mu_i - \mu_j$. Thus unrestricted direct optima are $v_i^* = \mu_i$ and $g_{ij}^* = \mu_i - \mu_j$. For Equation 17, writing $\delta_i = v_i - \mu_i$ leaves the excess conditional risk

$$\frac{1}{K}\sum_i \delta_i^2 + \frac{1}{P}\sum_{i<j}(\delta_i - \delta_j)^2.$$

Its unique minimum is $\delta = 0$. The absolute-only objective has the same population optimum. Conditional means lie in the allowed closed loss and difference ranges; finite sigmoid/tanh models may only approach boundary values. Under deterministic execution and unrestricted access to the full prompt and generator, q can itself be reconstructed, so this is no claim of irreducible execution uncertainty.

The boundary asserts neither finite-sample equality nor equality of restricted approximation classes. Different fitted encoders, bottlenecks, regularizers and optimization can yield different residuals even with the same raw evidence. The common-encoder arm isolates head fitting; end-to-end arms additionally allow representation changes. A comparison of heads with different masks or private context does not test this boundary. Direct predictions need not be cycle-consistent away from the optimum, whereas differences of scalar potentials always are. Such flexibility offers no additional target information and no universal learning or tail advantage.

L.7 A restricted error regime for direct comparison

Write $q_i = b + a_i$ and absolute prediction error as $v_i - q_i = e + \eta_i$. The shared residual e cancels exactly in $(v_i - v_j) - (q_i - q_j)$, leaving $\eta_i - \eta_j$. Across a declared anchor population, assume centered action-specific residuals with variance τ^2 and covariance $\rho\tau^2$, and centered direct-head residual $g_{ij} - (q_i - q_j) = \xi_{ij}$ with variance κ^2. Then

$$\mathbb{E}[(v_i - v_j - d_{ij})^2] = 2\tau^2(1-\rho), \qquad \mathbb{E}[(g_{ij} - d_{ij})^2] = \kappa^2.$$

Thus direct comparison has strictly lower pairwise MSE when $\kappa^2 < 2\tau^2(1-\rho)$. This describes a possible advantage when a shared scalar predictor leaves action-specific fitting errors that a pair-conditioned predictor can reduce. It is an explicit assumption on attained residuals, not a theorem

that the proposed architecture achieves such residuals or learns faster. If absolute errors are entirely common ($\eta_i = \eta_j$), subtraction is exact and no direct predictor can improve its MSE. Biases add their squared mean-difference terms and must be measured, not assumed away.

Neither this decomposition nor lower MSE establishes narrower simultaneous tail intervals. The comparison head must also achieve a favorable tail profile and satisfy the calibration-budget condition in Proposition 3. Report held-out residual covariance, bias, pairwise error and tail quantiles for direct and derived differences. Calibrating $v_i - v_j$ separately can exploit error cancellation even without a learned pair head, making it a substantive comparator rather than a redundant restatement of the absolute box.

L.8 Family dependence and distribution shifts

If several transforms or evidence banks share a source behavior, they cannot be counted as independent calibration anchors. Either choose one transform/bank by a predeclared random rule per independent source family, or define one score per family by maximizing Equation 15 over its entire fixed panel. The latter gives simultaneous coverage for an exchangeable new family panel of the same design. It does not imply coverage for arbitrary new attack families or for an attacker whose distribution changes in response to observed defenses. Run feedback-driven attacks as stress tests with their own empirical reporting; do not attach Proposition 1 to them without new assumptions.

The primary theorem uses exact adjudicated losses of deterministic executions. Substituting noisy Monte Carlo estimates would provide coverage for those estimates, not automatically for expected policy values. For a stochastic extension, construct simultaneous intervals for all calibration and test action values and propagate their uncertainty, with a separately budgeted failure probability. This extension is not implemented or claimed here. Likewise, annotation error relative to real-world harm is not corrected by conformal coverage: the target is the declared evaluator's loss.

L.9 Class-relative opportunity versus fitted-critic error

For nested selector families, $\Pi_h \subseteq \Pi_L$ after embedding short inputs by truncation, infimum monotonicity gives $V_L \leq V_h$. Adding and subtracting their infima yields Equation 20. This decomposition does not make either component empirically identifiable from one fitted model. Multiple restarts, capacity controls, and a separately trained action-only arm compare attainable procedures; they do not prove proximity to the family optimum. No universal semantic resolution limit follows. In particular, an unrestricted computation on a known prompt and deterministic model can reconstruct the greedy completion without a sampled future. The substantive question is the value of an explicit computational representation for a restricted, learned selector.

L.10 Complementary certificates and action gain

Table A17 contains $q = (0.04, 0.40)$ in both panels. It is representable with $c_J = 1$, absolute centers $(0.15, 0.30)$ and $(0.40, 0.60)$, absolute scales 0.15 and 0.40, and pair centers/scales $(-0.29, 0.31)$ and $(-0.38, 0.02)$. Single-view closure gives $V_1 = 0.30$ and 0.64; joint closure gives 0.09. Shifting both true losses and absolute intervals upward by 0.50 leaves differences unchanged but raises the joint bound to 0.59. Thus relative order cannot anchor graded acceptability. For binary harmful loss, however, $q_i \leq \beta < 1$ already entails zero regret; independent regret effects require graded losses.

Complete action-gain construction. For Table A18, set $K = 3$ and use the two reached horizons $\{4, 8\}$ with $L = 8$. Prescribe the reference priority $2 \prec 3 \prec 1$ and hence $a_0 = 2$. This isolates the selection algebra; it is not a simulation of the primary $K = 4, L = 64$ H16-backed experiment. With $q = (0.04, 0.30, 0.30)$, absolute errors at $h = 4$ are $(0.26, 0, 0.25)$, within $r^v = 0.30$; errors at $h = 8$ are $(0.46, 0.20, 0.20)$, within 0.50. True differences are $(-0.26, -0.26, 0)$; absolute pair errors at $h = 4$ are $(0.46, 0.46, 0.48)$ and those at $h = 8$ $(0.36, 0.50, 0.46)$, all at most 0.50. Both sets are nonempty and valid. A common rank quantile of one is stipulated, not estimated.

Antisymmetric raw row maxima are $(0.20, 0.48, 0)$ at $h = 4$ and $(0.10, 0, 0.76)$ at $h = 8$, so their unique choices are 3 and 2. Exact closure gives

views	(V_1, V_2, V_3)	(U_1, U_2, U_3)	Γ_{21}
$h = 4$	$(0.60, 0.60, 0.35)$	$(0.60, 0.60, 0.30)$	-0.60
$h = 8$	$(0.74, 1, 1)$	$(0.60, 0.40, 1)$	-0.60
both	$(0.09, 0.39, 0.35)$	$(0, 0.30, 0.30)$	0.24

The bound on z_3 at $h = 4$, combined with the $z_1 - z_3$ bound at $h = 8$, yields $V_1 \le 0.09$. Adding $z_3 - z_2 \le 0.02$ from $h = 4$ gives $z_1 - z_2 \le -0.24$ and $U_1 = 0$. The feasible vector $(0.09, 0.33, 0.35)$ attains $V_1 = 0.09$ and $z_2 - z_1 = 0.24$, proving equality. Both views are necessary for qualification and a positive Γ_{21}; without either, the fixed-panel deletion rule chooses its prescribed reference action 2. With both panels, action 1 is the only qualified candidate, so the priority rule selects it.

Direct upper bounds at $h = 4$ for $(12, 13, 23, 21, 31, 32)$ are $(0.70, 0.70, 0.98, 0.30, 0.30, 0.02)$, so no direct exclusion is available at margin ε. The only negative upper bound at $h = 8$ is $u_{13} = -0.26$, excluding 3, not 2; survivor ranking against the original candidate set still chooses 2. Closure at $h = 4$ also has no negative pair bound, so a propagated-veto control based on $h = 4$ does not explain the correction. Absolute-only cumulative bounds are $(0.60, 0.60, 0.35)$ and qualify nothing. These are fixed-output comparisons, not dominance claims after retraining or recalibration. Copying action 3 as a fourth candidate with identical losses/centers and $g_{34} = 0$ embeds the example in $K = 4$; action-key ties choose between the copies without changing the gain.

From local support to population accounting. Equation 10 follows by feasible-set minimization and remains valid after selecting the ordered pair. Qualification alone bounds regret by ε and loss by β on the coverage event. Subtracting these from a baseline with nonnegative loss/regret proves the regression bound in Section 4. Outside that event the loss difference can lie anywhere in $[-1, 1]$. A positive example or marginal coverage therefore does not establish positive average gain. Section 4.3 distinguishes certificates that need multiple views from changes that do; Appendix I retains all minimal subsets and the actual deletion outcomes.

M Execution Supervision and Calibration Protocol

The following settings define trace-wise prediction, execution labels, and calibration. Benchmark label banks use LLM-proxy judgments. The independent human-adjudication protocol is specified separately below; human and model-generated labels are distinct evaluation sources.

Execution labels. Freeze the harmful/benign objective and rubric before viewing responses. Harmful loss is one when the complete answer materially advances the stated harmful objective through usable assistance, and zero otherwise; a disclaimer does not negate actionable content. For benign requests, prewrite $m_s \ge 1$ observable binary criteria, including correctness. If b_{ir} records criterion satisfaction, use

$$q_i = 1 - m_s^{-1} \sum_{r=1}^{m_s} b_{ir},$$

with loss one for unjustified refusal or harmful redirection. Necessary clarification is evaluated by an explicit checklist criterion. Report refusal and task-quality endpoints separately. This rubric defines the evaluator's operational loss, not universal utility.

For human confirmation, two blinded annotators score each unique prompt/answer pair and a third resolves disagreements; randomize presentation and retain supporting spans. Reuse labels only for identical pairs. Freeze the resulting table before fitting or confirmation selection. Unresolved outcomes remain unknown. An unresolved calibration-family score can conservatively be $+\infty$; incomplete confirmation cannot support a complete endpoint. Proxy and human labels remain separately identified, with measured agreement and actual adjudication counts.

Each anchor stores its source family, split, prompt/prefix, rubric, ordered candidate token IDs, greedy completions, evidence seeds, model/tokenizer/backend revisions, annotations, q_i, d_{ij} and hashes. Labels supervise predictions and offline regret only; they are not deployment features.

Generation and frozen readout. The reference configuration uses a Qwen3.5-9B generator and a Qwen3.5-4B preference critic with the existing adapter and ranking head frozen. Set $K = B = 4$, $M = 2$, $\mathcal{H} = \{4, 8, 16, 32, 64\}$ and a 512-token total-answer cap. Draw 16 proposals with temperature 0.9, top-$p = 0.95$, top-$k = 50$, retaining four distinct openings; insufficient diversity triggers top-first-token branching with greedy completion of each opening. Futures use temperature 0.8, top-$p = 0.9$, top-$k = 50$; final execution is greedy.

Preserve the original serializer, random streams and per-call slot maps. Capture the terminal input to the critic's original output projection in one unpadded forward; no readout token is added. Reset attention/recurrent states between examples. Both heads receive the same $[r_h^{(m)}; e_i^{(m)}]$, see only the declared prefix, and share weights across traces/horizons. Restore candidate identities before averaging outputs and computing training/calibration residuals. Appendix J gives the head formulas and parameter count. Check unchanged preference logits, padding/readout equivalence, unseen-suffix isolation and packed-anchor isolation before model-level evaluation.

Optimization and selection. Train only the added heads on cached readouts with AdamW: learning rate 10^{-3}, betas $(0.9, 0.999)$, epsilon 10^{-8}, weight decay 0.01, gradient cap one, batch size 16 independent anchors, five epochs, 5% linear warm-up and cosine decay. Average both unit-weighted loss terms over all five horizons within each anchor. Select minimum development loss, breaking epoch ties toward the earlier checkpoint. Use seeds 17/29/43, with 17 primary and independent calibration per fit. Candidate permutation requires re-encoding and is a separate ablation. A separate end-to-end control may train rank-16, scale-32, zero-dropout LoRA; it is not the primary method.

After development selection, freeze RMS scales with floor 10^{-3}:

$$\sigma_h^v = \max\{10^{-3}, [\textstyle\sum_{s,i}(v_{s,i,h} - q_{s,i})^2/(N_d K)]^{1/2}\},$$

$$\sigma_h^g = \max\{10^{-3}, [\textstyle\sum_{s,i<j}(g_{s,ij,h} - d_{s,ij})^2/(N_d P)]^{1/2}\}.$$

Scales are common across candidates, with no test prompt-type input. The default design tolerances are $\alpha = 0.05$, $\beta = 0.10$, $\varepsilon = 0.05$. The head-diagnostic panel in Appendix F instead reports its declared $\alpha = 0.10$ setting; these calibration levels are kept distinct. These are study settings, not validated application-safety requirements. For 1,000 calibration families, the default $\alpha = 0.05$ rank is 951; at $\alpha = 0.10$ it is 901. Precision targets and source-family requirements are stated once in Appendix I.

Training and deployment algorithm. Split source families; freeze generation, execution and rubric; generate all candidate completions and labels; fit heads and select on development; freeze scales/tolerances; calibrate the entire horizon path on independent families. At deployment, form the five-view panel, retain the reference score/key ordering, intersect constraints, and compute exact closure. Choose the first jointly qualified candidate; an empty, unresolved or unqualified panel returns a_0 uncertified. Commit its B tokens and complete under the same greedy policy. Log all bounds, selected actions, conflicts, failures and generation/critic/closure/completion costs. Confirmation scores all candidate executions for regret and the delivered answer for benchmark endpoints.

Direct and derived controls. The absolute-only arm minimizes $\mathcal{L}_{\text{abs}} = \mathbb{E}_{s,h} K^{-1} \sum_i (v_i - q_i)^2$. The primary derived arm has no pair head and minimizes

$$\mathcal{L}_{\text{der}} = \mathbb{E}_{s,h}\left[K^{-1}\sum_i (v_i - q_i)^2 + P^{-1}\sum_{i<j}((v_i - v_j) - (q_i - q_j))^2\right]. \tag{17}$$

Pairwise gradients update both scalar outputs. A derived-after-absolute arm forms differences after absolute-only training. Match frozen readouts, data, seeds, selection budget, updates and information access; report parameter counts and a development-fixed capacity/compute match. Calibrate each fitted arm jointly over its own coordinates and full path; absolute-only omits pair coordinates. Separate nominal-α intervals do not imply joint error α.

Record residual bias/covariance, MSE, tail quantiles, widths, cycle errors, propagation margins and action outcomes. Derived cycles vanish algebraically; direct cycles may help fit or worsen coherent calibration. The fixed-interval component grid and independent deployment comparisons are in

Appendices F and I. Proposition 3 holds for fixed absolute predictions, not separately retrained arms. Variants changing action length, intervention position, generator, critic or evidence policy require their own declared training/calibration population and full-population evaluation. Nominal horizons and parameter counts do not establish latency savings.

N Reference Implementation and Extensions

Backbone compatibility. The original preference risk is $1 - p_i$; risks sum to $K - 1$, so they are not marginal harmfulness probabilities. Preserve SHA-derived action-key ties and the per-call slot maps. The reference order averages the two original trace scores. Newly supervised heads average outputs after restoring these same identities. The historical cyclic-order refinement in Appendix D is a separate comparator; rotations neither prove permutation invariance nor validate the new controller.

Failed scalar retrofit. An archived nonnegative affine calibration of relative scores collapsed to slope zero. Across 2,012 cached anchors it predicted approximately 0.00847868 everywhere, with zero qualified/changed actions at $\beta = 0.10, \varepsilon = 0.05$ and identical loss 0.00776590 for three selectors. This motivates investigating hidden readouts, without establishing their benefit.

Executable interface and checks. The package provides shared NumPy heads with analytic gradients, anchor-level calibration and exact rational closure. A read-only pre-hook captures the original output-projection input and returns the scorer's unmodified result. It runs once and is removed on failure; require evaluation mode, frozen parameters and an unpadded single example. The panel adapter calls the original builder/scorer for ten horizon–trace records, versus two for standalone H16; feature capture adds no backbone forward. The selector sees no realized labels and returns the original action when disabled or unqualified.

Implementation checks cover gradients, trace exchange, slot mapping, calibration, closure, hook cleanup, and critic call counts. Ranker compatibility requires unchanged original logits and candidate identities. Benchmark performance is evaluated separately through the executed outcomes.

Optional first-stop extension. The main rule evaluates all views at $L = 64$. A separate stopping arm checks $\{16, 32, 64\}$, including views 4/8 at the first check, and commits the first qualified action in the fixed reference order. Conflict or exhaustion returns uncertified a_0. Proposition 1 covers the same jointly calibrated path; H16 is available at each eligible stop. Post-stop full-panel comparisons are offline counterfactuals. Qualification and savings require measurement of all proposal, evidence, critic, closure and completion work.

O Additional Method Analysis

Absolute-derived differences can cancel shared prediction error, so a direct comparison head is not inherently superior. Appendix L.7 gives its conditional error advantage. The following result asks when propagation pays for a larger joint calibration radius, holding absolute predictions fixed.

Proposition 3 (a sufficient net certification gain). Hold the absolute predictions and development scales fixed. Let $c_A, c_J < \infty$ be the separately computed absolute-only and joint calibration quantiles, and suppose both cumulative sets are nonempty. In the unmerged constraint graph, write a path's cost at radius c as $\mu_p + c\lambda_p$, with $\lambda_p \geq 0$ the sum of its frozen scales. Let A_i be action i's absolute-only upper bound and $\Delta = (c_J - c_A)_+$. If a path $p : 0 \to i$ and paths $p_k : k \to i$ for all competitors satisfy

$$\mu_p + c_A\lambda_p + \Delta\lambda_p \leq \beta < A_i, \qquad \mu_{p_k} + c_J\lambda_{p_k} \leq \varepsilon \quad \forall k, \tag{18}$$

then i qualifies under the joint rule and fails the absolute-only loss gate. The propagation margin $\beta - (\mu_p + c_A\lambda_p)$ must cover the calibration-inflation cost $\Delta\lambda_p$. This separates useful relational structure from merely adding prediction coordinates. It is a checkable, action-specific sufficient condition; extra prompt coverage additionally requires that no absolute-only action qualified earlier or at this horizon. The comparison isolates calibration cost for the same absolute predictor, not superiority to a separately retrained model. Appendix L.5 gives the graph parameterization, proof and a constructed case with $c_J > c_A$.

O.1 What the horizon contrast can identify

For any frozen selector, retain the diagnostic

$$D_E(h, L; s) = q_{\hat{a}_h} - q_{\hat{a}_L} = R_E(h; s) - R_E(L; s). \tag{19}$$

It is a policy-value contrast, not a new theorem or an information-theoretic lower bound. To separate two possible explanations, let Π_h be a declared resource-restricted selector family and $V_h = \inf_{\pi \in \Pi_h} \mathbb{E} q_{\pi(F_h)}$. If Π_L contains every short-horizon selector by allowing prefix truncation, then $V_h - V_L \geq 0$. For a fitted selector in each family, write $A_h = \mathbb{E} q_{\hat{a}_h} - V_h \geq 0$. The observed contrast decomposes as

$$\mathbb{E} D_E(h, L) = (V_h - V_L) + (A_h - A_L). \tag{20}$$

The first term is opportunity from longer evidence *within the declared family*; the second is unequal fit to that family. Neither is identified by one horizon curve. Retrained capacity/seed controls probe the second term without claiming to recover the infimum. Because a fixed prompt and known deterministic generator already determine the execution, an unrestricted selector could in principle compute it from the prompt alone. We therefore make no claim of irreducible missing information. Nested families do not imply equal total compute; equal-cost comparisons are separate experiments.

P Supporting Diagnostic: Visibility and Reliability

This diagnostic motivates execution-aligned prediction and investigates a possible source of errors. It does not establish the necessity or effectiveness of cumulative inference. Fix openings, execution policy and full-answer losses q_i. Generate complete sampled branches once; let $b_i^{(m)} \in \{0, 1\}$ label each complete branch on harmful anchors. Reveal nested prefixes F_h without changing either outcome table. For branch predictions $\hat{b}_{i,h}^{(m)}$ and a frozen action selector $\hat{a}_h$, define $\mathcal{B}(h) = \mathbb{E}[(\hat{b}_{i,h}^{(m)} - b_i^{(m)})^2]$ and $\mathcal{R}(h) = \mathbb{E}[q_{\hat{a}_h} - \min_i q_i]$. The diagnostic signature is

$$\mathcal{B}(L) < \mathcal{B}(h), \qquad \mathcal{R}(L) > \mathcal{R}(h), \quad h < L. \tag{21}$$

Average traces/candidates within anchors and anchors within independent source families. Retain the full population, including initially correct, incorrect and tied choices. Improved Brier error alone is not decision improvement; worse selection without improved branch recognition is generic deterioration. If both improve, longer evidence helps both tasks. Unresolved differences are not equivalence; a plateau needs a predeclared margin.

A frozen-representation probe tests decodability, while the same-output audit scores unchanged action-head predictions against both execution and branch targets. Conditional suffix resampling and blinded targeted/sham edits test branch-specific reliance with q fixed; new branches receive new labels. These do not identify a unique neural mediator. Appendix Q specifies class-conditional calibration, family sampling and interventions; Appendix S gives an existence construction and no-deterioration cases. Because $F_h \subseteq F_L$, an unrestricted selector can ignore the additional suffix. The hypothesis concerns fitted evidence use, not information loss; separate long-horizon retraining and controls with informative additional evidence remain required.

Q Visibility Diagnostic Protocol

This supporting diagnostic protocol defines branch-level analyses separately from the main evaluation population. Freeze executed outcomes q and complete sampled-branch outcomes b before revealing prefixes. Evaluate the joint Brier/regret signature in Equation 21 on all source families, including initially wrong choices and ties. Average candidates/traces within anchors and anchors within families. Report class-conditional error, calibration and a counter-regime where longer evidence helps both recognition and decisions. A plateau requires a predeclared equivalence margin.

Same-output and representation controls. Score unchanged action-head predictions against both q and complete-branch labels. Fit a slot/trace-conditioned sigmoid linear branch probe on the frozen common readout; Appendix S specifies features and regularization. Probe decodability alone does not establish action-head use. Match training data, capacity and development selection for horizon-specific refits; if refitting removes the signature, report a fitting limitation.

Natural resampling. Keep each candidate's first eight tokens fixed and sample four complete suffix banks with seeds 101, 211, 307 and 401, using recorded independent candidate/trace streams. Already terminated branches remain terminated. Apply the original 64-token evidence and 512-token answer caps. Relabel all new complete branches blind to selector outcomes, keep every bank, and leave executed q fixed. Use all six bank pairs and all six action pairs, averaged by source family. Report action disagreement, regret range and the directional co-response $\Delta g_{ij}\Delta\bar{b}_{ij}$ with identical pair orientation. Label-discordance strata are descriptive, not randomized groups; selector-effect filtering is prohibited.

Blinded edits. Freeze edit eligibility before scoring. Construct branch-targeted and matched sham edits preserving the declared prefix and controlling length/style; blind editors and annotators to method predictions. Retain all eligible cases, failed edits and validity judgments, relabel complete edited branches and keep q fixed. Compare targeted versus sham responses with a separately corrected diagnostic family. Edits on selected cases support only that eligible population, and neither edits nor resampling uniquely identify neural mediation.

Opportunity and controls. On binary harmful losses, all-safe and all-unsafe banks have regret zero, with selected loss zero and one respectively. Thus

$$\mathbb{E}R_E(\hat{a}) = \Pr(\text{mixed})\,\Pr(q_{\hat{a}} = 1 \mid \text{mixed}).$$

The conditional term is undefined for an empty mixed stratum. Report stratum weights and unfiltered outcomes; deployment never uses outcome-defined strata. Keep benign graded losses separate. Match behavior-/execution-supervised heads, current/cumulative constraints, and removal of the absolute gate. The latter chooses the first reference-ranked action with $U \le \varepsilon$ or returns a_0, and provides no absolute-loss guarantee. Source populations, multiplicity and execution-consistent cost controls follow Appendix I. Recognition improvement plus deterioration supports this diagnostic signature; deterioration alone does not identify its cause.

R Supporting Analysis: What Constraint Retention Protects

This appendix analyzes a conditional method property, independent of whether the visibility–reliability diagnostic is observed. Exclusions from retained evidence may remain useful while unresolved choices change. For a nonempty feasible set at a reached horizon h, $\mathcal{Q}_{\le h}$, define a comparison witness and its excluded actions at margin $\delta \ge 0$:

$$W_{ij,h} = \max_{z \in \mathcal{Q}_{\le h}} (z_i - z_j), \qquad \mathcal{V}_h(\delta) = \{j : \exists i,\ W_{ij,h} < -\delta\}. \tag{22}$$

The existing closure computes $W_{ij,h} = D_{ji}$. A witness says that *every* loss vector compatible with $\mathcal{Q}_{\le h}$ ranks j worse than some i by more than δ; it does not require committing to i as the final winner.

Witness preservation lemma. If $j \in \mathcal{V}_h(\delta)$ and $\delta \ge \varepsilon$, no nonempty set at $h' \ge h$ obtained by retaining $\mathcal{Q}_{\le h}$ and intersecting additional constraints can qualify j under the regret gate. This remains algebraically true for arbitrary additional constraints. If the actual $q \in \mathcal{Q}_{\le h}$, then $R_E(j) > \delta$. Empty intersections disable certification; the fallback is not protected by this lemma.

The full proof, including compactness and a counterexample to robustness against arbitrary additional constraints, appears in Appendix S.

A measurable bridge to actual improvement. For fixed-panel attribution at L, let a_0 be the original H16 selector and a_C the cumulative qualified choice, with the identical a_0 fallback whenever it does not qualify. Then, exactly,

$$R_E(a_0) - R_E(a_C) = \mathbf{1}\{C\}\,[R_E(a_0) - R_E(a_C)]. \tag{23}$$

On the simultaneous coverage event, if C occurs and $a_0 \in \mathcal{V}_h(\delta)$ for $\delta > \varepsilon$, the gain exceeds $\delta - \varepsilon$. Thus the theory points to three observable quantities: witness availability at h, selection at L of actions excluded at h, and successful qualification of an alternative. None is guaranteed frequent. Log the full gain decomposition, including losses on the remaining qualified cases; the existence of a witness alone does not imply a population gain.

Minimal control. Appendix S specifies direct prefix veto and its propagated variant. They test whether retaining exclusions alone explains any gain; the full method must also earn its complexity against truncation and learned fusion.

S Restricted Model, Proofs and Diagnostics

S.1 Exact two-target construction

For two actions let $d_E = q_1 - q_2$ and $d_B = b_1 - b_2$. Suppose an initial estimate is $\hat{d}_0 = -m < 0$ and the branch comparison is opposed, $d_B > 0$. Under the explicitly assumed mixing model,

$$g_h = (1 - w_h)\hat{d}_0 + w_h d_B \quad \text{reverses sign iff} \quad w_h > \frac{m}{m + d_B}, \qquad 0 \le w_h \le 1, \tag{24}$$

If the initial ranking is correct, this change worsens execution; otherwise it can correct an error. The assumed mixture is not asserted to describe every neural critic. It is one sufficient construction for the main diagnostic pattern.

The construction summarized in Appendix P is an existence model of fitted behavior, not a claim that training on execution labels produces the stipulated weight. Let T_1, T_2 be independent $\mathrm{Bernoulli}(\rho)$ variables with $0 < \rho < 1/2$, let $q = (0, 1)$, and let $b = (T_1, 1 - T_2)$. At two ordered views $h_1 < h_2$, evidence at h_1 reveals q but not the flips; evidence at h_2 includes both flips. Thus the Bayes predictions at h_1 for branch labels are $(\rho, 1 - \rho)$ and their mean Brier error is $\rho(1 - \rho)$. Predictions at h_2 equal b, with zero error. All labels refer to complete branches and remain fixed across the horizon reveal.

The action comparison at h_1 is $g_0 = -1$ and selects action 1, with zero regret. Let $g_L = (1 - w)(-1) + w(b_1 - b_2)$. If both flips are zero, $g_L = -1$; if exactly one flips, $g_L = -(1 - w)$; if both flip, $g_L = 2w - 1$. For two actions the minimax rule with antisymmetric scores selects action 1 when $g_L < 0$ and action 2 when $g_L > 0$. Therefore its expected regret is ρ^2 for $1/2 < w < 1$ and zero for $w < 1/2$. At $w = 1/2$, double-flip cases tie; at $w = 1$, single-flip cases tie. Regret then depends on the declared token-key order, so no boundary-independent conclusion is asserted. An oracle using q chooses action 1 at either horizon, so the construction creates no information-theoretic loss.

For a general initial comparison $-m < 0$ and opposed branch comparison $d_B > 0$, the score is $-m + w(m + d_B)$. Its sign reverses exactly at $w = m/(m + d_B)$. If the initial execution ranking is wrong, reversal can improve the action; if d_B has the same sign, no reversal is induced by this convex mixture. These conditions motivate outcome-discordance strata and controls with informative additional evidence, not a claim that short futures dominate. A common additive shift of absolute scores leaves rankings unchanged; branch-specific *differential* error is required. The construction is therefore compatible with direct or absolute-derived differences and proves neither head superior.

S.2 Witness preservation under additional constraints

For finite calibrated widths, $\mathcal{Q}_{\le h}$ is a closed subset of the compact box $[0, 1]^K$. When nonempty, each maximum $W_{ij,h}$ is attained. If $W_{ij,h} < -\delta$, there is a strictly positive slack $\gamma = -\delta - W_{ij,h}$. For every z in any nonempty refined subset $\mathcal{Q}' \subseteq \mathcal{Q}_{\le h}$,

$$z_j - \min_k z_k \;\ge\; z_j - z_i \;\ge\; -W_{ij,h} \;= \delta + \gamma > \delta.$$

Consequently $\max_{z \in \mathcal{Q}'}(z_j - \min_k z_k) > \delta \ge \varepsilon$ and j cannot qualify. If $q \in \mathcal{Q}_{\le h}$, the same inequality shows $R_E(j) > \delta$. The graph formula $W_{ij,h} = D_{ji}$ follows from the already proved difference-constraint extrema in Appendix L. If widths are infinite, the full box has no such strict witnesses; no protection is claimed from absent evidence.

Set containment preserves exclusions. Interpreting them as true execution dominance additionally requires coverage of q. An inconsistent refinement can remove q while preserving an exclusion; an empty refinement triggers the reference fallback. The following example separates these cases.

For an explicit counterexample to overinterpreting robustness, let the initial set be $[0, 1]^3$ with $z_1 - z_2 \le -1/2$, and let the true $q = (0, 3/4, 1)$. Action 2 has an exclusion witness in that set. Additional

constraints $z_1 \geq 1/2$ and $z_3 \leq 0$ leave the nonempty singleton $(1/2, 1, 0)$, which falsely certifies action 3 at $\beta = \varepsilon = 1/4$. Action 2 remains excluded, but the true vector has been removed. This demonstrates why preserving a retained exclusion is weaker than valid full certification under arbitrary suffix corruption.

S.3 From qualification to realized gain

For fixed-panel comparisons with an identical fallback, $a_C = a_0$ whenever C fails, proving Equation 23 without a probabilistic assumption. On the simultaneous coverage event $\mathcal{E}$, a qualified cumulative action obeys $R_E(a_C) \leq \varepsilon$. If $a_0 \in \mathcal{V}_h(\delta)$ with $\delta > \varepsilon$, validity at h gives $R_E(a_0) > \delta$, hence the gain exceeds $\delta - \varepsilon$. On other qualifying cases in $\mathcal{E}$, the gain can be negative, but is at least $-\varepsilon$. Off $\mathcal{E}$, neither statement is guaranteed. Do not estimate $\Pr(\mathcal{E})$ from qualification alone or interpret the marginal coverage level as a conditional success probability.

Report the empirical identity as disjoint cells: qualification/nonqualification, legacy action excluded/not excluded at the declared witness margin, and positive/zero/negative realized regret change. For diagnostics, the fully adjudicated q table permits a separate audit of whether each retained set contains q; these offline labels never enter the deployed controller. The $h = 8$ witness is available before all eligible stopping checks, which begin at 16. For actual first stopping, count only witnesses computed no later than the decision and distinguish stopping before L from correcting an already observed reversal.

S.4 A reproducible minimal prefix-veto control

At the declared witness horizon $h = 8$, store all direct upper comparison bounds $u_{ij,t} = g_{ij,t} + r_t^g$ at reached $t \leq h$. Exclude j if some $u_{ij,t} < -\varepsilon$. Select at L the first survivor in the original H16 score/key order. If no survivor exists, report a control failure and use the original H16 fallback. Under a nonempty common feasible set and valid bounds, a true minimizer cannot be directly vetoed; empirical inconsistencies remain explicit.

Attribution uses the joint model's fixed intervals; deployment comparisons recalibrate on independent calibration data. The propagation-only variant replaces direct comparisons with closure witnesses through $h = 8$, preserving the survivor rule. Executed outcomes and qualification are evaluated separately for veto, full composition, long-only, and fusion.

S.5 Probe specification and interpretation

The branch probe is sigmoid-linear on $[r_h^{(m)} \otimes e_i^{(m)} \otimes e_m; e_i^{(m)}; e_m; e_h]$. It minimizes training-family Brier loss, with $L_2 \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ selected on development data. The critic remains frozen and sees neither unseen suffixes nor labels. Appendix Q defines resampling and blinded-edit comparisons; probe decodability and action-head use are separate quantities.

T Decision Boundaries and Worked Examples

Complementary evidence. Table A18 gives a mathematical example with losses $(0.04, 0.30, 0.30)$ and reference action 2. Intersecting both views yields $V_1 = 0.09$ and $U_1 = 0$, qualifying action 1 with improvement 0.26 and certified lower bound 0.24. Neither view alone qualifies an action.

Conflicting views. Intervals $[0, 0.012]$ and $[0.031, 0.079]$ for the same action have empty intersection. The controller therefore disables qualification and returns the fixed H16 action a_0. This is the fallback represented in the model diagram.

Pairwise evidence and priority. A bound on $q_D - q_B$ can establish B's advantage over D. Selection still requires $V_B \leq \beta$, $U_B \leq \varepsilon$, and reference-rank priority among qualified actions. The comparator bound Γ measures improvement over a_0.