ACTIONWISE SUPPORT CONTROL CLOSES THE GLOBAL-VETO GAP IN LOGGED TEST-TIME COMPUTE

Anonymous authors
Paper under double-blind review

ABSTRACT

Conservative logged controllers often decide whether to continue a language-model trajectory by collapsing actionwise support to a single state-level test. That aggregation can reject every continuation when one action is weakly covered, even when another action is supported and improves on a complete fallback. We introduce Actionwise Support Control (ASC), a feasibility operator that preserves the identity of supported actions until value comparison. ASC is defined for any fixed score, support statistic, fallback, and deterministic tie rule; it therefore composes with pessimistic estimators instead of replacing them. Our central result gives an exact gate-equivalence partition: an actionwise gate and a global gate disagree precisely on states with at least one failed support test and at least one support-eligible, fallback-improving action. The resulting global-veto mass is an interpretable diagnostic of where support aggregation can alter a deployed decision. We also give an outcome-isolated audit contract and a matched-comparison accounting contract that make empirical tests of this mechanism reproducible. Together, these results establish support topology as a first-class design choice for logged test-time compute: conservatism can discard unsupported actions without discarding qualified alternatives.

1 INTRODUCTION

Test-time allocation chooses whether a deployed language model should commit an answer or purchase another block of reasoning under a hard budget. In a single-trajectory deployment, that decision is consequential: the controller cannot replace a poor stop-or-continue choice by drawing a fresh set of answers. Conservative logged control is therefore attractive, but it has a structural choice that is usually left implicit. A state can have strong evidence for one continuation length and weak evidence for another. A controller may either filter those actions separately or compress their support to a single global veto.

This paper formalizes the cost of that compression. When a global gate rejects a state because one bundle is weakly covered, it can suppress another bundle that is both supported and preferred to the fallback. We call the set of such states the *global-veto set*. Its probability mass is a decision-level quantity: it identifies exactly where the aggregation rule, rather than the value estimator or action grid, can change the action.

Actionwise Support Control (ASC) keeps the support vector intact. Given the same scores, support statistic, threshold, fallback, and tie order, ASC discards only a failed action and compares the remaining eligible actions with fallback. The contrasting global gate returns fallback whenever any action fails the support test. This construction yields a sharp, estimator-agnostic result: the two gates agree everywhere outside the global-veto set and disagree at every state inside it. The result makes support topology auditable and supplies a direct design rule for conservative controllers.

The paper makes three contributions:

- A score-agnostic formulation of support topology for logged compute allocation, including the global-veto mass and a fully specified actionwise feasibility operator.
- An exact gate-equivalence partition that isolates the sole decision states on which global and actionwise conservative gates can differ.

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.

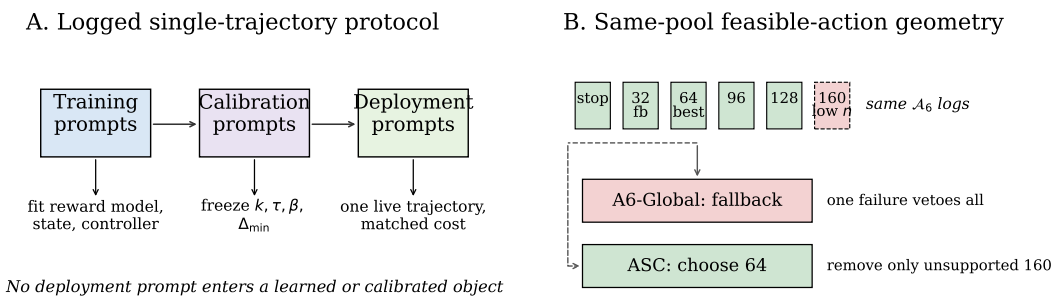

Figure 1: Support topology in a logged single-trajectory controller. A global gate rejects the full non-fallback action set when any action fails its support test. ASC removes only the failed action, then compares the remaining eligible actions with the complete fallback. The diagram is a mechanism schematic; its labels and line styles redundantly identify every decision path.

- Two reproducibility-oriented evaluation contracts: one for paired local outcome audits and one for matched cost and comparator accounting.

2 Related Work

Adaptive inference and search decide how much computation to allocate through repeated sampling, confidence-based exits, self-refinement, verifier-guided search, and learned continuation decisions (Snell et al., 2024; Brown et al., 2024; Schuster et al., 2022; Madaan et al., 2023; Shinn et al., 2023; Lightman et al., 2023; Yao et al., 2023; Wang et al., 2023; Ringel et al., 2025; Kim et al., 2026; Dong et al., 2026). These methods improve scoring, routing, or search. ASC addresses a distinct layer: whether the feasibility operator preserves the identity of actions long enough for an already chosen scoring rule to compare them.

Offline decision making provides doubly robust estimation, counterfactual risk minimization, high-confidence evaluation, pessimistic learning, and safe improvement under logged feedback (Precup et al., 2000; Swaminathan & Joachims, 2015; Thomas et al., 2015; Jiang & Li, 2016; Thomas & Brunskill, 2016; Jin et al., 2021; Rashidinejad et al., 2022; Kumar et al., 2020; Laroche et al., 2019). Those tools motivate conservative scoring. ASC is complementary: it specifies how action-dependent support constraints are aggregated after a score has been fixed. Selective prediction and abstention similarly calibrate rejection under uncertainty (Tayebati et al., 2025; Dong & Shinnou, 2026); a continuation controller additionally faces several actions with different support at the same state.

3 Logged Single-Trajectory Control

Consider a frozen reasoner evaluated at decision state s. Let $b(s)$ denote a complete fallback policy that advances the trajectory to answer commitment or a hard budget cap. The available non-fallback actions form a finite set $\mathcal{A}^+(s)$, for example stop and a set of feasible continuation bundles. The complete decision set is $\mathcal{A}(s) = \mathcal{A}^+(s) \cup \{b(s)\}$.

The formulation requires two quantities that are fixed before the gate is applied. First, $q(s, a)$ is a scalar score for every $a \in \mathcal{A}(s)$; it may be a pessimistic logged estimate, a learned value, or another pre-specified criterion. Second, $n(s, a) \geq 0$ is an action-specific support statistic for every $a \in \mathcal{A}^+(s)$. A logged implementation can compute n from known propensities and local neighborhoods, but the structural result below does not depend on a particular estimator. The fallback is always executable and is not support-gated.

For a fixed support threshold τ and a fixed fallback-relative margin $\delta > 0$, define

$$\mathcal{F}_\tau(s) = \{a \in \mathcal{A}^+(s) : n(s, a) \geq \tau\}, \qquad \mathcal{E}_{\tau,\delta}(s) = \{a \in \mathcal{F}_\tau(s) : q(s, a) - q(s, b(s)) \geq \delta\}. \quad (1)$$

For every nonempty set E, $\mathrm{sel}_s(E)$ is the $q(s, \cdot)$ maximizer in E, with ties resolved by a fixed total order $\prec$ chosen before any evaluation. This tie rule makes the policies below single-valued.

4 ACTIONWISE SUPPORT CONTROL

The global gate requires every non-fallback action to satisfy the support test before it permits a learned continuation. ASC filters actions separately:

$$\pi_{\mathrm{global}}(s) = \begin{cases} \mathrm{sel}_s(\mathcal{E}_{\tau,\delta}(s)), & \mathcal{F}_\tau(s) = \mathcal{A}^+(s) \text{ and } \mathcal{E}_{\tau,\delta}(s) \neq \varnothing, \\ b(s), & \text{otherwise}, \end{cases} \tag{2}$$

and

$$\pi_{\mathrm{ASC}}(s) = \begin{cases} \mathrm{sel}_s(\mathcal{E}_{\tau,\delta}(s)), & \mathcal{E}_{\tau,\delta}(s) \neq \varnothing, \\ b(s), & \text{otherwise}. \end{cases} \tag{3}$$

Thus both policies share every score, support statistic, threshold, fallback, feasible action, and tie rule. Their only distinction is whether one failed support test removes the entire non-fallback set.

For a reference distribution ρ over decision states, define

$$\mathcal{G}_{\tau,\delta} = \{s : \mathcal{F}_\tau(s) \neq \mathcal{A}^+(s),\ \mathcal{E}_{\tau,\delta}(s) \neq \varnothing\}, \qquad \Gamma_{\tau,\delta}(\rho) = \Pr_{s\sim\rho}[s \in \mathcal{G}_{\tau,\delta}]. \tag{4}$$

The global-veto mass $\Gamma_{\tau,\delta}(\rho)$ is the fraction of decision states at which global aggregation can reject a qualified action.

Proposition 1 (Gate-equivalence partition)**.** *For fixed q, n, τ, δ, fallback, feasible actions, and tie order, $\pi_{\mathrm{global}}(s) = \pi_{\mathrm{ASC}}(s)$ for every $s \notin \mathcal{G}_{\tau,\delta}$. For every $s \in \mathcal{G}_{\tau,\delta}$, $\pi_{\mathrm{global}}(s) = b(s)$ and $\pi_{\mathrm{ASC}}(s) \in \mathcal{E}_{\tau,\delta}(s)$.*

Proof. If $\mathcal{F}_\tau(s) = \mathcal{A}^+(s)$, Equations 2 and 3 apply the same selected set and the same tie rule. If $\mathcal{F}_\tau(s) \neq \mathcal{A}^+(s)$ and $\mathcal{E}_{\tau,\delta}(s) = \varnothing$, both policies return $b(s)$. These two cases cover the complement of $\mathcal{G}_{\tau,\delta}$. In the remaining case, the global rule returns $b(s)$ because its all-action support condition fails, whereas ASC selects the nonempty eligible set. □

The proposition is deliberately modular. Changing a critic, support statistic, or bundle grid may change the size of $\mathcal{G}_{\tau,\delta}$, but it cannot change the partition once those inputs are fixed. This separates an action-aggregation question from questions about estimator quality and search.

5 AUDITABLE EVALUATION CONTRACT

The theorem identifies which paired decisions are informative; an empirical audit must additionally prevent fitted estimates from being mistaken for outcomes. The following contract turns that distinction into a reproducible protocol.

5.1 PAIRED LOCAL-OUTCOME AUDIT

First select the audit states using a rule that is fixed without access to paired outcomes, and record the reference distribution, outer sampling unit, and every selected state identifier. At each selected state, materialize both the ASC-selected action and the complete fallback. Complete the two trajectories with the same decoding rule; when decoding is stochastic, pair the random streams by repetition. Record the number of paired completions, the terminal utility, the token and wall-clock budget, and every exclusion before aggregating outcomes.

The fitted score, support statistic, thresholds, and tie order must be frozen before these paired outcomes are examined. Fitting, calibration, state selection, and outcome measurement use disjoint named records. A reported local effect is therefore a paired outcome contrast on the selected states, not a reinterpretation of the score that selected them. When an interval is reported, its resampling unit is the prompt cluster; the report names the resampling method, number of resamples, confidence construction, and the exact set of simultaneous comparisons it covers.

5.2 Controller disclosure

An empirical ASC instantiation must make q and n reproducible. It reports the logged return target, reward-model architecture and fit split, propensity formula, neighborhood/radius rule, weighting and variance formula, cross-fitting folds when used, calibration objective and candidate set, selected hyperparameters, deterministic tie order, and every source of randomness. This disclosure makes the generic policies in Equations 2–3 operational without conflating the gate with a particular critic.

5.3 Matched comparisons and cost accounting

A comparison attributes a difference to support aggregation only when the competing gates share the feasible action grid, logs, score construction, support threshold, fallback, calibration records, random seeds, and budget target. Comparisons to search or multi-trajectory methods additionally disclose search permissions, verification access, candidate counts, decoding settings, and selection rules. Cost is measured with one declared unit and includes the same control and scoring overhead for every method. A report at equal accuracy identifies an observed operating point or states an explicit interpolation rule; it never treats an unreported point as matched. Grouped uncertainty names the grouping map and resamples those groups before object-level observations.

6 Discussion

ASC exposes a useful distinction in conservative deployment. Pessimism determines how a controller values an action; support aggregation determines whether that action reaches the comparison at all. The global-veto mass measures the part of a deployment distribution on which this second choice is active. The exact partition also sharpens experimental interpretation: a same-pool paired contrast can be attributed to the feasibility operator only when every input named in the contract is held fixed.

This perspective applies beyond a particular language model, estimator, or continuation grid. Any logged decision problem with multiple actions, action-specific evidence, and a safe fallback can report its support topology, veto mass, and paired decision audit. The result is a transparent route to conservative control that retains qualified alternatives while continuing to reject unsupported ones.

7 Conclusion

Actionwise Support Control makes support aggregation explicit in logged test-time compute. Its exact gate-equivalence partition identifies the only states on which a global veto can discard a qualified continuation, and the global-veto mass makes that mechanism measurable under a chosen deployment distribution. The accompanying paired-audit and matched-comparison contracts turn the distinction into an evaluable scientific claim. Conservative controllers can therefore preserve the benefits of support filtering while keeping supported, fallback-improving actions available for decision.

AI Use Statement

Generative AI was used only for manuscript writing, editing, and presentation. The authors reviewed the edits and remain responsible for the final content.

References

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024. URL https://arxiv.org/abs/2407.21787.

Sijin Dong and Hiroyuki Shinnou. Uncertainty-aware abstention in large language models with provable alignment guarantees. *arXiv preprint arXiv:2607.04430*, 2026. URL https://arxiv.org/abs/2607.04430.

Zhe Dong, Fang Qin, and Manish Shah. When does learning to stop help? a cost-aware study of early exits in reasoning models. *arXiv preprint arXiv:2606.30852*, 2026. URL https://arxiv.org/abs/2606.30852.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 652–661, 2016. URL https://proceedings.mlr.press/v48/jiang16.html.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5084–5096, 2021. URL https://proceedings.mlr.press/v139/jin21e.html.

Taehyeon Kim, Hyunsoo Lee, Youngsoo Jang, and Moontae Lee. Efficiently learning to reason or not to reason: Root-token policy optimization for adaptive thinking. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17934–17949, 2026. doi: 10.18653/v1/2026.acl-long.816.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 1179–1191, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0d2b2061826a5df3221116a5085a6052-Abstract.html.

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

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. URL https://arxiv.org/abs/2305.20050.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2023. doi: 10.52202/075280-2019.

Doina Precup, Richard S Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*, pp. 759–766. Morgan Kaufmann, 2000. URL https://web.eecs.umich.edu/~baveja/Papers/OffPolicy.pdf.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *IEEE Transactions on Information Theory*, 68(12):8156–8196, 2022. doi: 10.1109/TIT.2022.3185139.

Liran Ringel, Elad Tolochinsky, and Yaniv Romano. Learning a continue-thinking token for enhanced test-time scaling. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 3324–3345, 2025. doi: 10.18653/v1/2025.ijcnlp-long.177.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1269.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2023. doi: 10.52202/075280-0377.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. URL https://arxiv.org/abs/2408.03314.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 814–823, 2015. URL https://proceedings.mlr.press/v37/swaminathan15.html.

Sina Tayebati, Divake Kumar, Nastaran Darabi, Dinithi Jayasuriya, Ranganath Krishnan, and Amit Ranjan Trivedi. Learning conformal abstention policies for adaptive risk management in large language and vision-language models. *arXiv preprint arXiv:2502.06884*, 2025. URL https://arxiv.org/abs/2502.06884.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2139–2148, 2016. URL https://proceedings.mlr.press/v48/thomasa16.html.

Philip S Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015. doi: 10.1609/aaai.v29i1.9541.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://iclr.cc/virtual/2023/poster/11718.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2023. doi: 10.52202/075280-0517.

A Empirical Reporting Checklist

Table 1 provides a compact record structure for an empirical ASC study. It is designed to preserve the distinction between the formal partition and an observed performance contrast.

Table 1: Required record for an empirical ASC evaluation. Each field is fixed or disclosed before the associated claim is interpreted.

Record	Required content
Decision definition	Feasible action grid, complete fallback, score q, support statistic n, τ, δ, and fixed tie order.
Fit and calibration	Training/calibration split identities, return target, model and variance construction, folds, candidate set, calibration objective, and selected settings.
Paired audit	State-selection rule, action/fallback completion procedure, repetitions, shared randomness, terminal utility, exclusions, and prompt-cluster denominator.
Uncertainty	Outer resampling unit, group map when applicable, resample count, interval construction, and simultaneous-comparison scope.
Cost comparison	Declared cost unit, overhead inclusion, budgets, search and verification permissions, candidate counts, selection rule, and any interpolation rule.