

Beating a Random-Direction Bank Does Not Establish Policy Improvement: An Exact-Value Audit of Activation Steering for Agents

Avinash Goutham Aluguvelly
The University of Texas at Austin
algo@utexas.edu

Abstract

We audit whether outperforming random activation directions establishes improved agent performance. Using Llama-3.1-8B in a molecular identification task, we compute every legal action’s exact probability of eventual success under optimal continuation. An entropy-weighted steering direction outperforms all 39 layer- and norm-matched random controls on 42 held-out molecular pools in a frozen comparison. Post-hoc audits reveal the limits of this result. An untuned policy that chooses uniformly among legal actions also outperforms every random control, with no detectable difference from the steered policy; uncertainty intervals still admit modest effects in either direction. The steering direction ranks 10th among 36 directions when controls are matched to its construction, and its original-bank rank falls to 3/40 when pools sharing chemical scaffolds are grouped and weighted equally. Executed episodes show no detectable improvement over the unsteered baseline, with both policies performing near random under constrained decoding. These findings concern one model and task; they do not establish that steering cannot improve a competent agent. They demonstrate that superiority to a random-direction bank does not by itself establish policy improvement. We propose five reporting controls covering trivial baselines, direction construction, inferential units, state selection and horizon, and executed behavior.

1 Introduction

Activation steering changes model behavior without touching weights (Turner et al., 2023; Rinsky et al., 2024; Zou et al., 2023), and has been proposed for agent-facing uses such as modulating exploration (Rahn et al., 2024) and distilling procedural memory from past episodes (Zhao et al., 2026). A common test of whether a direction is real is a finite-reference comparison: sample many random vectors matched on layer and norm, deliver

them the same way, and show the learned direction beats them.

That comparison establishes less than it appears to. It shows that the direction does *something*, but it rolls five separate questions into one answer: whether behavior changed, whether task value improved, whether the construction rather than the resulting orientation is responsible, whether the effect holds up under the correct unit of inference, and whether any of it survives letting the model act. Recent work raises related worries. Xu et al. (2026a) separate target-concept preference (the steered output’s tendency toward the target concept) from coherent, task-valid generation, and Goyal and Daumé III (2026) separate having an effect from having the intended one. Xu et al. (2026b) find that steerability varies sharply across target behaviors, Ostermann et al. (2026) that steering does not yet substitute for weight updates, and Korznikov et al. (2025) that random directions can produce comparable behavioral change. What is missing is a setting where the true value of every action is known, so “the score went up” can be taken apart instead of taken on trust.

We build such an environment. An agent identifies a hidden molecule by asking binary structural questions, and a dynamic program gives each legal action its exact oracle-continuation value. On it we ran a steering study to its natural stopping point and then continued past it. A regret-derived direction failed two prospectively frozen tests. An entropy-weighted, EAST-inspired (Entropic Activation STEering) direction (Rahn et al., 2024) then ranked 1/40 against a hash-frozen bank of 39 norm-matched isotropic directions on 42 fresh pools. This is exactly the criterion a finite-reference comparison is designed to test. Four audits then show what the rank leaves open, and at the closed-loop endpoint we detect no agent-level benefit.

Is this an agent task? It is a deliberately small one: single-token action selection over 16 can-

didates on a four-question horizon, under constrained decoding. Sixteen candidates match the four-question budget exactly: $\log_2 16 = 4$ is the fewest binary questions that can distinguish 16 candidates under a perfectly balanced split, so there is no slack for a wasted question, though the structural predicates rarely split candidates that evenly in practice (Appendix A). That cost buys what the audit needs. Trajectory-level agent benchmarks (Yao et al., 2024) cannot supply the counterfactual value of every action a policy did not take, so there “steering improved the agent” can be measured but not decomposed; here it can be attributed or refused. The logic is that of a counterexample: one case where the rank holds and the improvement is not established shows the inference is not licensed, not that steering never helps.

Our contributions: (i) an exact-value constrained-action environment in which every legal action has a known counterfactual value; (ii) an audit that separates behavioral change, absolute value, finite-reference rank, trivial-policy superiority, construction specificity, independent-unit generalization, depth robustness, executability, and trajectory benefit into distinct claims with distinct controls (Table 1); and (iii) a case study in which the rank rung is met, everything above it is not, and a legal-uniform policy meets the same rung untuned. This is an evaluation audit and an agent-level negative result, not a steering method.

2 An Environment With Exact Action Values

Task and oracle. Each pool holds 16 molecular profiles built from ChEBI (Hastings et al., 2016) and RDKit structural predicates (RDKit contributors, 2026). An action asks for one unobserved binary property; evaluation states expose 50–53 legal actions. Under a uniform posterior over surviving candidates, $Q_s(a)$ is the exact probability of identifying the target within the remaining budget when a is forced first and an adaptive oracle plays optimally after that (Appendix A). This is the oracle’s value for continuing, not the value of what the model would do next. No prompt contains Q , the optimal action, or the target.

Estimand. Let $p_s(a)$ be full-vocabulary next-token mass on the single-token label of action a , $L_s = \sum_{a \in A_s} p_s(a)$ the legal mass, and $r_s = \max_a Q_s(a) - \min_a Q_s(a)$. Masking to legal la-

bels and renormalizing,

$$\bar{\pi}_s(a) = p_s(a)/L_s, \quad V_s(\bar{\pi}) = \sum_{a \in A_s} \bar{\pi}_s(a) Q_s(a), \quad (1)$$

$$\Delta \tilde{V}_s = [V_s(\bar{\pi}_{\text{int}}) - V_s(\bar{\pi}_{\text{base}})]/r_s. \quad (2)$$

This normalization, and the averaging of rotations within each pool followed by averaging pools that reports it, is the one written into the 42-pool reserve’s precommit, locally content-hashed on 2026-07-29 at 13:03 UTC and recorded status `frozen_before_sealed_reserve_model_calls`; it precedes model calls on the sealed reserve, not necessarily the opening of reserve pool identities and states, and it is a locally timestamped and hash-bound plan, not an externally registered preregistration (§C). Because L_s can be tiny, $\bar{\pi}_s$ tells us which action the model prefers once it is forced to choose one, not what it would emit on its own. We close that gap in §4. Since option symbols alone can move multiple-choice decisions (Yang et al., 2025), each state is scored under four fixed cyclic label rotations, averaged within pool before a two-sided sign-flip test or pool bootstrap. Because legal mass can fall to a median of 4.1×10^{-6} , $\bar{\pi}_s$ is a ratio of small quantities: logits are cast to float32 before the full-vocabulary softmax, the masked renormalization is recomputed in float64 with max-subtraction, and the two agree to 10^{-9} on every cell. All forwards ran on Llama-3.1-8B (Grattafiori et al., 2024) through NNsight/NDIF (Fiotto-Kaufman et al., 2025).

A two-candidate example. Near the end of a game, two candidates and one question remain: five of fifty legal questions separate them ($Q = 1$), the rest do not ($Q = 0$). Random scores .10, perfect scores 1.00; unsteered scores .1549 and steered .1570 (prompt excerpt in Appendix A; the complete prompt is in the released artifact) — the unsteered baseline is 5.49 percentage points above uniform, and steering adds a further increment of only .0021, the paper in miniature.

Selection, disclosed. Two stages pick the evaluation states, and neither looks at model output. The first keeps, for each pool and remaining budget, the state whose action values are furthest apart; the second takes the lowest state ID among a pool’s four. Because the first stage reads exact Q , the states we analyze are not typical task states but the highest-spread ones, and they fall unevenly across budgets:

Evidence level	Question	Result
Behavioral effect	Do legal-action preferences change?	Yes
Absolute value	Does exact value improve?	Inconclusive
Isotropic-bank rank	Does it beat a norm-matched bank?	Yes, 1/40
Trivial baseline	Does it beat legal uniform?	Not established
Construction specificity	Is the entropy construction special?	Not established, 10/36
Independent unit	Does it survive scaffold-level inference?	Not established, 3/40
Depth robustness	Does it survive depth balancing?	Not established
Label robustness	Is it invariant to action labels?	No
Executability	Does unrestricted decoding emit legal actions?	Only with no history
Trajectory benefit	Does terminal success improve?	None detected; near random

Table 1: The claims a finite-reference rank is often asked to carry. Each is a distinct question with its own control. The first row reports an observed change in legal-action preferences in this experiment, not a universal property of any intervention that changes next-token mass, since a change can fall entirely on illegal tokens or scale every legal token’s mass proportionally and leave the normalized legal-action policy unchanged. Of the substantive claims, only the third (the finite-reference rank) is met here, and an untuned uniform policy meets it too.

21, 14, 5, and 2 pools at budgets 1–4. The prospective rank below was therefore obtained on a favorable slice of the state space, not on arbitrary task states. We therefore also report a depth-balanced design in which all 42 pools contribute at every budget (§4).

3 The Result That Looks Like a Win

Directions. Decision regret is $\max_b Q_s(b) - Q_s(a)$. Contrast pairs share pool, target, turn count, terminal outcome, and query dictionary, differing only in whether the action history is oracle-optimal or lower- Q ; residuals are read at the final input token. The regret direction averages those differences. The EAST-inspired direction instead weights final-token residuals by masked action entropy H_{pi} and centers within two-state packs, $\mu_p = (h_{p1} + h_{p2})/2$:

$$d_H = \frac{\sum_{p=1}^{12} \sum_{i=1}^2 H_{pi}(h_{pi} - \mu_p)}{\sum_{p=1}^{12} \sum_{i=1}^2 H_{pi}}. \quad (3)$$

This is inspired by EAST (Rahn et al., 2024) rather than a reproduction: it replaces sampled completions and run-level centering with the masked next-token distribution and pack centering, and intervenes at one prompt token (Appendix B). Layers {8, 16, 24} and coefficients {−.10, −.03, .03, .10} were calibrated on source data only.

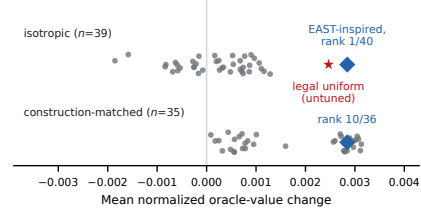


Figure 1: The same 42-pool effect against two reference banks. The frozen isotropic bank matches layer, L2 norm, and delivered RMS; the construction-matched bank additionally shares the source pipeline, its occupied subspace, or the activation covariance. The direction is extreme in the first and unremarkable in the second. An untuned legal-uniform policy also beats every isotropic control.

Two failed gates, then a rank of 1/40. On a 12-pool holdout the regret direction gave +.000918 ($p = .0835$) with reverse steering at −.000888. The signs were coherent, but the gate failed because two of eight matched isotropic directions did better. On 50 fresh pools it fell to +.000331 ($p = .2727$), fifth of 40. EAST-inspired was a prespecified secondary there, and its advantage motivated one further question: would it outrank isotropic directions delivered at its own larger RMS? Before opening a pre-existing 42-pool reserve we hash-froze EAST-inspired, its reverse, 39 new Gaussian controls, all prompts, and all rotations. The reference statistic s_{ref} is the plus-one exceedance fraction over the 40-direction bank (Phipson and Smyth, 2010), treated as descriptive.

No control tied or exceeded it. EAST-inspired ranked **1/40**, the minimum attainable $s_{\text{ref}} = .025$, with a .001560 margin, stable under every one-pool deletion. Absolute improvement was weaker: +.00285, 26/42 pools positive, CI [−.000530, .006049], frozen $p = .1017$. Reverse steering was also positive (+.00085) and forward-minus-reverse inconclusive ($p = .226$). Under the finite-reference criterion alone, this is a passing result.

4 Audits

A: an untuned policy clears the same bar. The endpoint is a distribution over legal actions, so the legal-uniform policy $\bar{\pi}_s(a) = 1/|A_s|$ is admissible and carries no learned content. It gains +.00247, or 87% of the EAST-inspired effect, and independently exceeds all 39 isotropic controls (Figure 1). Head to head we detect no difference ($p = .949$ normalized, $p = .927$ raw), uniform ahead in 25/42

pools under both, on intervals wide enough to admit modest effects either way. On raw value uniform is the larger of the two (Table 5); both rank 1/40 unnormalized, so the $1/r_s$ weighting in Eq. 2 is not doing this. A bank matched only on layer, norm, and RMS does not separate this learned direction from an untuned flattening control.

B: matching construction, not just norm. An isotropic vector shares none of the structure our direction inherits from its source data, so we built 35 controls that do: permuted entropy weights, per-pack sign flips, leave-pack-out rebuilds, random directions inside the source subspace, and Gaussians carrying the source activation covariance. We ran them on identical states, rotations, layer, and delivered RMS (Appendix D; 6,384 forwards). Rebuilding d_H from stored activations reproduced the frozen tensor hash exactly, and re-running the baseline reproduced the earlier pool means, so the two banks are comparable. Against these references the direction ranks **10/36**. Excluding near-copies *raises* it (1/14 among controls with $|c| \leq .9$), because the directions that beat it are precisely those sharing its orientation. Since every control is delivered at the same RMS, one at cosine c to d_H places a component $.10c$ along that axis, and pool-mean effect is near-linear in c ($r = .90$, $R^2 = .81$), with d_H sitting 0.21 residual SD above the fitted line (App. D). So the safe reading is a narrow one: the isotropic bank identifies an effective *orientation*, but cannot attribute that orientation to the entropy-weighted construction.

C: the inferential unit and the horizon. Pools were checked for candidate-ID disjointness, which does not imply chemical independence. Connecting pools that share an exact nonempty non-chiral Bemis–Murcko scaffold (Bemis and Murcko, 1996) and taking the transitive closure collapses the 42 pools into 27 components, and 25/42 share a scaffold with an earlier experimental role. Weighting the 27 components equally, the unit an independent-family claim implies, drops the effect to $+.00139$ ($p = .391$) and the rank to 3/40, below the best random control. Horizon matters as much: the selection rule puts 35/42 analyzed states at a remaining budget of at most two. Scoring all four budget states for all 42 pools shrinks the effect, and its pool-level $p = .023$ does not survive this paper’s own scaffold unit, where it is $+.000906$, CI $[-.00032, .00218]$, $p = .173$ (App. E). Depth balancing removes about a third of the one-state-per-

pool effect, and neither estimate is distinguishable from zero.

D: the executed endpoint. A one-step score hands the rest of the episode to an oracle, so it cannot answer the agent question. We therefore executed the episodes (App. F), applying the intervention at every step. This evaluates the masked-greedy realization rather than the full masked distribution scored above, so it tests deployment relevance rather than the trajectory value of an identical stochastic policy. Because outcomes are binary over a four-question horizon, expected terminal identification probability is computed *exactly* by enumerating realized branches rather than by sampling, and is comparable to the oracle’s V . Over 504 episodes (3,420 nodes) the oracle attains 1.000 and a uniform random questioner .0678; the steered policy reaches .0651 and the unsteered baseline .0580. We detect no benefit from steering (steered minus baseline): $+.00707$ by pool ($p = .226$), $+.00606$ by component (95% CI $[-.00481, .01693]$, $p = .295$). Neither policy separates from random: EAST-inspired minus uniform is $-.00273$ by pool ($p = .605$) and $-.00006$ by component ($p = .992$). The baseline is below random by pool ($p = .030$) but not by component ($p = .218$), so we report that as suggestive only. The episodes ran at all only because decoding was constrained: the unrestricted top token was legal at 504/504 first steps and 0/2916 thereafter.

Mechanism and labels. The direction did not reliably raise its own entropy measure on either set ($p = .298$ on 50 pools, $p = .147$ on 42). Label sensitivity is its own limit on the endpoint: 36/42 pools produced both positive and negative effects across the four rotations, and one rotation alone would have ranked 14/40. The aggregate is stable because it averages over a nuisance dimension in which the underlying quantity is not. In this setup legality depends sharply on prompt state: the unrestricted top token is legal in all 168 no-history cells, where the median absolute intervention effect is 3.6×10^{-8} , and in none of the 504 history-present cells. The one regime in which this model could emit a legal first action unaided is the one in which the direction does nothing.

5 Exploratory: Calibration Transport

Repeating the identical construction and grid on Llama-3.1-70B and 405B (Grattafiori et al., 2024)

and Olmo-3-32B (Allen Institute for AI, 2026) changed the outcome every time: across the three Llama scales the selected relative depth moved 25%, 50%, 75% and the coefficient reversed at 405B, of 24 transported operating points 20 fail the receiving model’s own criterion, and where the uniform comparison was reached it is again null. These runs used mutable deployments and pools we had already opened, so we report them in App. G as descriptive transport checks rather than as evidence that the finding generalizes.

6 What to Report

Exactly one claim here is prospective: a rank of 1/40 in one frozen isotropic bank at $s_{\text{ref}} = .025$, and on a favorable slice of states. Everything after is post hoc, diagnosing why that rank is insufficient rather than falsifying it. We apply that discipline to our own analyses: the scaffold reaggregation is a sensitivity check on the candidate-ID split’s independence assumption, not a demonstration that Bemis–Murcko components are themselves statistically independent or the uniquely correct unit. We co-report pool-weighted and equal-component estimates throughout as a post-hoc sensitivity analysis, fixed regardless of outcome; equal-pool and equal-component aggregation weight observations differently and so are not interchangeable standard errors for one estimand. Two pool-level findings cross $p < .05$; neither does under equal-component aggregation over the 27 components. We therefore regard those findings as sensitive to aggregation. Before a finite-reference rank is read as agent improvement, we suggest reporting five controls, each of which moved our conclusion here: (1) **trivial-baseline**, a trivial untuned policy under the same endpoint; (2) **construction-matched**, a control bank matched on *construction*, not only norm; (3) **inferential-unit**, the unit of inference the claim needs, not the one the split gives; (4) **selection-and-horizon**, the selection rule and the horizon it samples; and (5) **executed-endpoint**, an executed endpoint with its legality rate. Without exact Q , (1) and (5) need only the deployed policy and still apply unchanged; (2)-(4) still apply against a task-appropriate held-out sample and the task’s own disjointness structure (e.g. near-duplicate items or shared templates), in place of a frozen reserve and its exact per-item value.

Limitations

Nothing here establishes that steering cannot help a competent agent. Legality is forced throughout this evaluation, and both policies sit near a random floor far below an oracle that solves every pool, so what we bound is what this class of evidence supports, not the ceiling of activation steering. The primary audit is one base 8B model, one synthetic molecular environment, one compact state representation; the cross-model ladder of App. G is exploratory and does not extend it. That ladder reaches the uniform comparison on only two further models, covers no isotropic bank, construction-matched bank, depth-balanced design, or closed-loop episodes at any scale, and therefore yields no finite-reference rank outside 8B, so nothing in it is comparable to the 1/40 this paper audits. Its deployments were served at mutable labels, its pools were already inspected, and one operating point rests on a tie-break between two equally close depth roundings. $Q_s(a)$ assumes oracle-optimal continuation after the forced first action, not the model’s future behavior. The four rotations are cyclic shifts of a single label family, so they expose label sensitivity without addressing label-token geometry; independent permutations or label-free pointer scoring would be stronger. Equal Euclidean perturbations need not have equal behavioral effect (Nguyen et al., 2026). The construction-matched bank spans the source pipeline, its subspace, and the activation covariance but not the output-head subspace, and holds layer and coefficient fixed, so it bounds specificity at one operating point. Every audit was designed after the reserve outcome was visible; each control bank was hash-frozen before its forwards and reported regardless of outcome, but a post-hoc bank is not a preregistered one. The scaffold audit re-partitions an already-frozen sample; a prospectively scaffold-disjoint split remains the right fix and is future work, as is an instruction-tuned model that can emit the action vocabulary unaided, on which the closed-loop conclusion could differ. The closed-loop runner records the executed action and terminal outcome but not the legal-action logits it chose from, so state and prompt reconstruction and terminal values are independently replayable while whether each action was the model’s true masked argmax is not; future runs should retain those logits. Trajectory-level agent benchmarks are broader than this task (Yao et al., 2024). The NDIF deployment label was mutable despite frozen local artifacts and

canary checks.

Ethics Statement

This work uses ChEBI (Hastings et al., 2016), an open chemical ontology, and RDKit-derived structural predicates (RDKit contributors, 2026). No human subjects, personal data, or annotator labor were involved, and no dual-use synthesis information is produced: the environment asks coarse structural yes/no questions about already public molecules. All model calls were single-token forward passes on shared hosted models. The primary audit used 29,548 scientific forwards and 47 deployment canaries on one 8B model across seven completed studies, including both failed prospective gates, plus 756 calls in an earlier calibration session halted by an infrastructure fault. The exploratory ladder of App. G adds 5,448 scientific forwards and 111 canaries across four further models, two of which stopped before any comparison. We report both separately so audit cost is not read as evidence of scale, and so the exploratory total is not read as part of the primary study. Activation steering is dual-use, and our finding argues for narrower claims: a direction can win a norm-matched comparison while showing no detectable advantage over a uniform policy, so steering results should not be treated as reliable capability or safety interventions without the controls above. A complete, offline-replay-verified artifact (prompts, configurations, frozen vectors, per-cell outcomes, both failed gates, the cross-model ladder) is prepared and hosted at <https://github.com/tokenledger/steering-policy-audit> (private at the time of writing; **to be made public before the camera-ready deadline**).

References

- Allen Institute for AI. 2026. Olmo 3. <https://allenai.org/olmo>. Open language model release.
- Guy W. Bemis and Mark A. Murcko. 1996. **The properties of known drugs. 1. molecular frameworks.** *Journal of Medicinal Chemistry*, 39(15):2887–2893.
- Jaden Fiotto-Kaufman, Alexander R. Loftus, Eric Todd, Jannik Brinkmann, Koyena Pal, Dmitrii Troitskii, Michael Ripa, Adam Belfki, Can Rager, Caden Juang, Aaron Mueller, Samuel Marks, Arnab Sen Sharma, Francesca Lucchetti, Nikhil Prakash, Carla Brodley, Arjun Guha, Jonathan Bell, Byron C. Wallace, and David Bau. 2025. **NNsight and NDIF: Democratizing access to open-weight foundation model internals.** In *The Thirteenth International Conference on Learning Representations*, pages 92337–92370.
- Navita Goyal and Hal Daumé III. 2026. **Steering safely or off a cliff? rethinking specificity and robustness in inference-time interventions.** In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. **The Llama 3 herd of models.** *arXiv preprint arXiv:2407.21783*.
- Janna Hastings, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck. 2016. **ChEBI in 2016: Improved services and an expanding collection of metabolites.** *Nucleic Acids Research*, 44(D1):D1214–D1219.
- Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Y. Rogov, Ivan Oseledets, and Elena Tutubalina. 2025. **The rogue scalpel: Activation steering compromises LLM safety.** *arXiv preprint arXiv:2509.22067*.
- Tam Nguyen, Tu Anh Nguyen, Sina Alemohammad, and Richard G. Baraniuk. 2026. **Minimizing collateral damage in activation steering.** *arXiv preprint arXiv:2605.01167*.
- Simon Ostermann, Daniil Gurgurov, Tanja Baeumel, Michael A. Hedderich, Sebastian Lapuschkin, Wojciech Samek, and Vera Schmitt. 2026. **From weights to activations: Is steering the next frontier of adaptation?** In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29854–29879, San Diego, California, United States. Association for Computational Linguistics.
- Belinda Phipson and Gordon K. Smyth. 2010. **Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn.** *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 39.
- Nate Rahn, Pierluca D’Oro, and Marc G. Bellemare. 2024. **Controlling large language model agents with entropic activation steering.** *arXiv preprint arXiv:2406.00244*.
- RDKit contributors. 2026. **RDKit: Open-source cheminformatics software.** <https://www.rdkit.org>. Software, version 2026.03.4; archive at <https://doi.org/10.5281/zenodo.591637>.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. **Steering Llama 2 via contrastive activation addition.** In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522. Association for Computational Linguistics.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. **Steering language models with activation engineering.** *arXiv preprint arXiv:2308.10248*.
- Ziwen Xu, Chenyan Wu, Hengyu Sun, Haiwen Hong, Mengru Wang, Yunzhi Yao, Longtao Huang, Hui Xue, Shumin Deng, Zhixuan Chu, Huajun Chen, and Ningyu Zhang. 2026a. **Why steering works: Toward a unified view of language model parameter dynamics.** In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31719–31736. Association for Computational Linguistics.
- Ziwen Xu, Kewei Xu, Haoming Xu, Haiwen Hong, Longtao Huang, Hui Xue, Ningyu Zhang, Yongliang Shen, Guozhou Zheng, Huajun Chen, and Shumin Deng. 2026b. **How controllable are large language models? a unified evaluation across behavioral granularities.** In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31269–31299, San Diego, California, United States. Association for Computational Linguistics.
- Zhen Yang, Ping Jian, and Chengzhi Li. 2025. **Option symbol matters: Investigating and mitigating multiple-choice option symbol bias of large language models.** In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1902–1917. Association for Computational Linguistics.
- Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. **τ -bench: A benchmark for tool-agent-user interaction in real-world domains.** *arXiv preprint arXiv:2406.12045*.
- Chengfeng Zhao, Yuqiao Tan, Shizhu He, Yequan Wang, Jun Zhao, and Kang Liu. 2026. **Neural procedural memory: Empowering LLM agents with implicit activation steering.** *arXiv preprint arXiv:2606.29824*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. [Representation engineering: A top-down approach to ai transparency](#). *arXiv preprint arXiv:2310.01405*.

A Oracle and Task Details

Let S be the surviving candidates, A the unanswered binary properties, and b the remaining budget. For outcome $o \in \{0, 1\}$, $S_{a,o}$ is the subset consistent with asking a and observing o , and $\Omega(S, a) = \{o : S_{a,o} \neq \emptyset\}$. Set $K(S, A, b) = |S|$ when $|S| \leq 1$, and zero for $|S| > 1$ when $b = 0$ or no action splits S . Otherwise

$$K(S, A, b) = \max_{a \in A} \sum_{o \in \Omega(S, a)} K(S_{a,o}, A \setminus \{a\}, b - 1), \quad (4)$$

and under a uniform posterior $V(S, A, b) = K(S, A, b)/|S|$ with

$$Q(S, A, b, a) = \frac{1}{|S|} \sum_{o \in \Omega(S, a)} K(S_{a,o}, A \setminus \{a\}, b - 1). \quad (5)$$

The implementation enumerates only realized nonempty branches and retains ties as multiple optimal actions. Each pool has 16 candidates and 53 binary RDKit-derived properties; the 42 reserve states expose 50, 51, 52, or 53 legal questions in 21, 14, 5, and 2 pools, with within-state Q ranges 1, .75, .625, and .5625 at the same frequencies.

A worked example. The task is a guessing game. We pick one molecule in secret out of sixteen. The model is shown the candidates still consistent with what has been asked so far, the questions already asked and their answers, how many questions remain, and a menu of the yes/no questions it is still allowed to ask. Every question on the menu is tagged with a single letter. The model’s entire job is to reply with one letter: the question it would ask next. It never answers a question, it never sees our secret, and it is never told which question is best.

Here is a state near the end of a game. Two candidates are left and one question remains, so the model either names a question that tells them apart and wins, or it does not and loses. Abridged, it is shown:

```
Candidate cards (identities are anonymous):
- M01: {"structure":
  "COc1cc(CC[NH3+])cc(OC)c1OC"}
- M05: {"structure":
  "NC(=O)OCC(COC(N)=O)c1cccc1"}
Current survivors: M01, M05
Observation history:
- Does the molecule have substructure
  phenol? Answer: no
- Is the molecule’s rotatable at least 4?
  Answer: yes
- Is the molecule’s logp at least 2?
  Answer: no
```

```
Remaining question budget: 1
Remaining legal questions:
- A: Is the molecule’s fraction csp3 at
  least 0.75?
- B: Is the molecule’s logp at least 0?
- C: Is the molecule’s molwt at least
  350?
- D: Is the molecule’s rings at least 2?
  ... 44 further labeled questions ...
- E: Is the molecule’s hba at least 4?
- W: Does the molecule have substructure
  ester?
Choose the question that maximizes the
probability of identifying the molecule
within the full remaining budget.
Do not answer a question and do not
explain your choice.
Return exactly one action label from:
A, B, C, ..., W, x
Action label:
```

Of the 50 questions on the menu, exactly five tell M01 and M05 apart: is hba ≥ 4 (E), is hbd ≥ 2 (Z), is heteroatoms ≥ 6 (q), is there an amide (m), and is there an ester (W). Asking any of those five identifies the molecule with certainty, so each is worth 1. The other 45 leave both candidates alive, so each is worth 0. On this state, then, a policy’s score is simply the chance it names one of the five.

Naming a letter uniformly at random scores $5/50 = .10$. A perfect player scores 1.00. The unsteered model scores .1549 and the steered model .1570: both a little better than guessing, both nowhere near perfect, and barely different from each other. That is the paper in miniature.

One more thing this state makes visible. Those scores are computed after discarding everything the model would rather have said. Across its whole vocabulary, only 4.2×10^{-6} of its next-token probability lands on any of the 50 letters at all; we renormalize what is left. So the quantity being scored is which question the model prefers *when forced to pick one*, not what it would do if left alone.

B Direction Construction

Each regret contrast places the same decision state and label rotation after two equal-token-length curriculum contexts that share pool, hidden target, turn count, terminal outcome, and public query dictionary but differ in action and survivor history: the good trajectory selects oracle-optimal questions, the regret trajectory deterministically prefers feasible informative actions in the bottom half of strictly positive-regret actions while preserving the matched terminal outcome. With 24 states and four rotations, $d_R^\ell = \frac{1}{24} \sum_s \frac{1}{4} \sum_r (h_{sr}^{+, \ell} - h_{sr}^{-, \ell})$, normalized to unit RMS before the signed coefficient. Source-only calibration selected layer 8

(coefficients below); the negative sign was empirical, and reverse interventions use the opposite sign.

The EAST-inspired direction is Eq. 3. With two states per pack each numerator term is proportional to $(H_{p1} - H_{p2})(h_{p1} - h_{p2})/2$, so

$$d_H \propto \sum_{p=1}^{12} c_p (h_{p1} - h_{p2}), \quad c_p = \frac{1}{2}(H_{p1} - H_{p2}). \quad (6)$$

Two consequences follow from stored activations with no model calls. First, replacing the entropy weights by a constant makes d_H identically zero, since the two centered contributions within a pack are exact negatives, so an “unweighted centered” variant of this construction does not exist, and a unit-contrast control (the signs of c_p with magnitudes discarded) replaces it. Second, the contrast is concentrated: the shares of $\sum_p |c_p|$ are .246, .245, .244, and .241 for four packs and at most .006 for each of the remaining eight, the smallest 4.6×10^{-7} . Four of twelve packs therefore account for 97.5% of the absolute entropy-contrast weight $\sum_p |c_p|$; because d_H sums $c_p(h_{p1} - h_{p2})$, coefficient shares alone do not establish shares of the resulting vector, since hidden-state difference norms and cancellation across packs also matter. We rebuilt d_H from stored residuals and entropies and recovered the frozen tensor hash bit-exactly before relying on either property.

Calibration, selection, and canaries, disclosed.

The 24 direction-construction states comprise two fresh states from each of 12 curriculum packs (MCP01–MCP12, frozen 2026-07-27T19:06:48Z, [configs/ndif_semantic_symmetry_v2_precommit.json](#)). The source partition excluded 44 previously inspected states and selected construction states by deterministic SHA256 ranking within each pack, choosing the two minimum-ranked fresh states for direction construction and one further fresh state each for tuning and internal holdout; this data-partition rule is distinct from, and does not use, any performance-based ranking. A separate, performance-based operating-point selection stage, described below, ranks settings on the 12-*pool* calibration set and is not the rule that chose which states entered d_H .

Final 8B calibration used a separate 12-*pool* set under the value-and-entropy calibration plan frozen on 28 July 2026 at 11:19:38 UTC ([configs/ndif_8b_value_entropy_calibration_v1.json](#)), which replaced and

supersedes an earlier regret-only diagnostic frozen 2026-07-28T10:22:36Z ([configs/ndif_8b_source_calibration_v1.json](#)) that was aborted before completion; we retain that earlier plan only in the historical accounting and do not use it for final selection. After averaging the four rotations within each pool, regret settings were ranked by maximizing median normalized-value change, then mean change, smaller absolute strength, and earlier layer; EAST-inspired settings were ranked by the analogous rule on normalized-entropy change. Each method additionally required a positive mean and median change in its own objective and at least 8 of 12 positive pools; EAST-inspired further required mean legal-mass change of at least -0.02 . This selected layer 8 with coefficients $-.03$ for regret and $-.10$ for EAST-inspired. Every scientific run in this study ran a deployment canary at the start and end of each uninterrupted process, on the first frozen state/rotation at the study’s layer, gated at a maximum absolute legal-logit change of 0.02 and a maximum relative residual-RMS change of 0.005; a process whose recorded canary exceeded either tolerance was not merged into the checkpoint it was running.

Molecular data come from ChEBI ontology version `chebi/253` (retrieved 2026-07-28, CC BY 4.0) with predicates computed under RDKit 2026.03.4. The hosted model was `meta-llama/Llama-3.1-8B`, recorded revision `d04e592bb4f6`; the runner requested the mutable NDIF label `main`, so recorded revision and canaries bound the weights only up to that label. Prompts were plain text, not chat templates. Each isotropic control was drawn iid Gaussian at a fresh frozen seed, L2-matched, unit-RMS normalized, and delivered at the target coefficient; those controls match neither the activation covariance nor the construction pipeline. Label rotations are deterministic cyclic assignments with offsets derived from SHA-256 identifiers, and all four are averaged before pool-level inference.

C Study Status

Stage	Status	Use
Source construction	development	built both directions
Calibration	development	layer and coefficient per method
12-pool holdout	prospective	first regret gate; failed
50-pool validation	prospective	regret replication; failed; EAST-inspired was a frozen secondary
42-pool reserve	prospective follow-up	EAST-inspired finite-bank criterion; 1/40
Uniform, temperature, scaffold, budget controls	post hoc	stored logits and exact Q
Construction-matched bank	post hoc	35 controls; rank 10/36 (App. D)
Depth-balanced states	post hoc	all four budgets per pool (App. E)
Closed-loop episodes	post hoc	trajectory endpoint (App. F)
Sealed confirmation set	sealed	not scored or claimed

Table 2: “Frozen” means locally content-hashed before outcomes were opened; it is not an external preregistration.

The 12-pool test enumerated all sign assignments. The 42- and 50-pool frozen runners used two-sided deterministic 100,000-draw Monte Carlo sign-flip tests with plus-one correction, and the frozen interval used 20,000 pool-bootstrap draws. Paper-facing post-hoc intervals use 100,000 pool resamples and sign-flip tests use 1,000,000 draws at fixed recorded seeds. All scientific cells, failed gates, per-rotation outcomes, rendered prompts, vectors, frozen configurations, and canary results are retained.

D Construction-Matched Controls

Post hoc: the reserve was opened before these controls were designed, so this is not a prospective gate. Its plan was hash-frozen before its forwards and it was reported regardless of outcome. It reuses the frozen states, rotations, prompt hashes, layer, coefficient, and delivered RMS byte-for-byte; only the bank changes. It committed 6,384 forwards over 38 conditions with the same canary tolerances and exact-resume checkpointing as the frozen study, and all five canaries stayed within tolerance. Baseline and d_H were re-run rather than reused, and

their pool means reproduced the earlier values to 0 difference with pool-level $r = 1.000$, so the two banks are directly comparable.

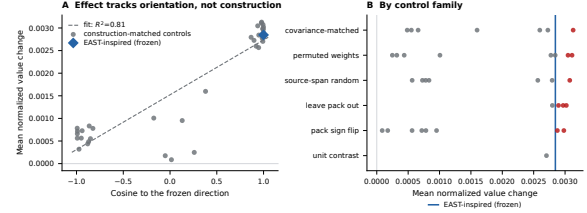


Figure 2: **A:** every control is delivered at the same layer and RMS, so one at cosine c to the frozen direction places a component $.10c$ along that axis; pool-mean effect is close to linear in c and the frozen direction (diamond) lies within a quarter of a residual SD of it. **B:** the same effects by family; red points exceed the frozen direction. Nine of 35 controls do.

The 35 scored controls are one unit-contrast direction, four informative leave-pack-out rebuilds, eight permuted-entropy-weight directions, eight per-pack sign flips, seven random directions in the span of the pack contrasts, and seven Gaussians carrying the sample covariance of the 96 stored source residuals; random families redraw on any $|\cos| \geq .9995$. Cosines to d_H span $-.993$ to $+.999$. The frozen direction ranks 10/36, beaten by a covariance-matched Gaussian ($+.003129$), a permuted-weight direction ($+.003111$), a source-span random direction ($+.003075$), a second permuted-weight direction, three leave-pack-out rebuilds, and two per-pack sign flips. Per family it ranks 2/8 against covariance-matched, 2/8 against source-span, 3/9 against permuted weights, 3/9 against sign flips, and 4/5 against leave-pack-out. Near-antiparallel controls ($c \leq -.90$) average $+.00062$ and none exceed it, consistent with reverse steering also being weakly positive in the frozen study. Because four packs carry 97.5% of $\sum_p |c_p|$, seven leave-pack-out rebuilds are collinear with d_H at $|\cos| \geq .9995$; after unit-RMS normalization their delivered perturbations differ by far less than the .02 canary tolerance, so they are empirically indistinguishable interventions rather than literally the same one, and they were predicted rather than remeasured. The pack with the smallest contrast was retained as a positive control and reproduced the frozen effect to 0 difference.

E Depth-Balanced Evaluation

Post hoc, reported regardless of outcome. The frozen study scored one state per pool, leaving 21, 14, 5, and 2 pools at budgets 1–4, so its deeper strata were uninterpretable. This study scores all four budget-specific states for all 42 pools (168 states, 672 cells, 2,016 forwards over baseline, EAST-inspired, and reverse; uniform and the oracle bound come exactly from stored Q). Of the 168 states 126 had never been scored; the 42 overlapping the frozen study reuse its rotation namespace, so their prompts are byte-identical and reproduced the earlier values to 0 difference. Prompt length grows with remaining budget (median 975.5, 1080.5, 1213.5, and 1521 tokens), and five budget-4 states exceed the frozen study’s self-imposed 2048-token ceiling, the longest at 2362 tokens; excluding them would have reintroduced a depth-correlated selection, so the ceiling was raised after confirming the longest prompt completes on the same deployment.

Per budget the effect is $+0.003053$ ($p = .054$), $+0.000374$ ($p = .803$), $+0.001616$ ($p = .169$), and -0.000000 ($p = .963$), each on 42 pools. The paired within-pool contrast between budgets 1–2 and 3–4 is $+0.000905$ ($p = .394$), which does not support the reversal the frozen five- and two-pool strata suggested. Pooling depths equally gives $+0.001261$ ($p = .023$) by pool but $+0.000906$, CI $[-0.00032, .00218]$, $p = .173$ over the 27 components, with per-budget component figures $+0.002250$, -0.000435 , $+0.001810$, and -0.000000 . Against the one-state-per-pool component effect of $+0.00139$, depth balancing removes about a third and neither estimate is distinguishable from zero. Against legal uniform the per-budget differences are -0.004872 , -0.000493 , $+0.000762$, and $+0.000046$, none distinguishable from zero.

Legality splits perfectly on whether a state carries an observation history. At budget 4, where there is no history and all 16 candidates are alive, the top token is legal in 168/168 cells and total legal mass has median .687; at budgets 1–3 it is legal in 0/504 cells and legal mass has median 4.1×10^{-6} . The frozen study’s eight legal cells were exactly its budget-4 cells. The intervention’s median absolute effect is 8.9×10^{-3} , 7.4×10^{-3} , and 8.9×10^{-3} at budgets 1–3 but 3.6×10^{-8} at budget 4, where baseline mass on the optimal set is 4.5×10^{-8} and the model’s preferences lie inside a Q -equivalent class.

F Closed-Loop Episodes

Post hoc; reported regardless of outcome, and the only study here whose endpoint is a trajectory outcome. An episode begins at a pool’s full-budget state with all 16 candidates alive. At each step the policy scores the legal action labels, the highest-scoring label is executed, the environment answers it against the hidden target, survivors are filtered, and the budget decrements; the intervention is applied at every step. The endpoint is terminal identification probability under the uniform posterior, which is what the oracle maximizes, so values are directly comparable to the frozen V . Because outcomes are binary over a four-question horizon, this expectation is computed exactly as $W(S, A, b) = \sum_o (|S_{ao}|/|S|) W(S_{ao}, A \setminus a, b-1)$ with $a = \pi(S, A, b)$, so a deterministic policy has an exact value after at most $2^4 - 1$ forwards per episode and carries no Monte Carlo error. Substituting the oracle for the model reproduces the stored root V exactly, which verified the harness before any forwards were spent. Note that this evaluates the masked-*greedy* realization of the model whereas the one-step endpoint evaluates its full masked *distribution*; the two are different policies, so this tests deployment relevance rather than extending an identical policy to a trajectory.

The study ran 42 pools $\times$ 4 rotations $\times$ 3 conditions: 504 episodes and 3,420 decision nodes, averaging 6.8 nodes rather than the maximum 15, because the model’s questions frequently fail to split the survivor set. The oracle attains 1.000, a uniform random questioner .0678, EAST-inspired .0651, its reverse .0662, and the unsteered baseline .0580. Steering gives $+0.00707$ by pool (95% CI $[-0.0034, .0175]$, $p = .226$) and $+0.00606$ by component (95% CI $[-0.00481, .01693]$, $p = .295$). EAST-inspired minus uniform is -0.00273 by pool ($p = .605$) and -0.00006 by component ($p = .992$). The baseline is below random by pool (-0.00980 , 95% CI $[-0.0181, -.0013]$, $p = .030$) but not by component (-0.00612 , 95% CI $[-0.0153, .0034]$, $p = .218$), so we treat it as suggestive only. As a share of the oracle-minus-random span, baseline closes $-.011$ and EAST-inspired $-.003$. Across the 3,420 nodes the unrestricted top token was legal at 504/504 first steps and 0/2916 later steps, with median legal mass falling from .687 to about 4×10^{-6} once any observation entered the prompt.

G Cross-Model Calibration Ladder

Exploratory throughout. Every non-8B deployment was served at a mutable label with no immutable revision, and the 42 pools were already open, so these are descriptive transport checks and replications on inspected data. Each model rebuilt its own source activations, direction, and operating point from scratch: nothing frozen transfers, because the residual width differs at every scale. Depth-matched grids place candidate layers at 25/50/75% of each model’s own depth, since a wrong-depth index is otherwise accepted silently. All figures below are computed from stored artifacts with no model calls.

Model	G0	G1-R	G1-H	Depth	α
Llama-3.1-8B	~ .97	fail	pass	25%	-.10
Llama-3.1-70B	.9101	fail	pass	50%	-.10
Llama-3.1-405B	.9685	pass	pass	75%	+.10
Olmo-3-32B	.9334	pass	fail	—	—

Table 3: The cross-model ladder. G0 is the minimum split-half reconstruction cosine of d_H over the three depth-matched layers, gated at $\geq .90$ and frozen before the statistic existed; G1-R and G1-H are the regret and entropy calibration branches, each judged on its own native objective. Depth and α are the entropy branch’s selection. Olmo-3-32B earned no operating point: its entropy branch missed the legal-mass floor by .0014 while clearing the other three criteria at 12/12 pools. Four other Olmo variants would have passed, but the frozen rule selects first and gates second, so switching after seeing the failure is the retuning the precommit forbids.

Transport. The grids share a design, so an inherited (relative depth, α) pair always names a cell that exists in the target’s grid and whose statistics are already stored. Evaluating each model’s selection against every other model’s frozen criterion gives 24 ordered pairs over both branches, of which **20 fail** and none reproduces the target’s own selection. On the entropy branch alone, 9 of 12 fail. Three pass (8B→70B, 70B→Olmo, and Olmo→70B), so the honest claim is that inherited operating points usually but not always fail, and never coincide with what the target model would have chosen. We report this rather than the stronger universal version because the data do not support the latter.

Two decouplings. At 405B the regret branch passes its calibration criterion at 9/12 pools while its direction reconstructs at split-half cosine .036, essentially unreconstructed across source halves. The two gates measure different properties, so this is not a contradiction, but it means a task-facing

calibration gate can pass on a direction that is not reproducibly estimated, and a researcher reporting only the calibration outcome would present that operating point as sound. Separately, the 405B entropy branch clears its native criterion with a mean exact-value delta of $-.001135$: the proxy the direction was constructed to move improves while the exact-value point estimate does not. Because we give only the point estimate here, we claim only that, and not that the direction reduces value.

The comparison, at the paper’s own unit. Re-computed over the 27 scaffold components, the equal-component estimate §4 co-reports alongside the pool level as a fixed, outcome-independent sensitivity check:

Model	mean	median	p	$+/-n$
Llama-3.1-8B	+.003102	-.008777	.733	10/27
Llama-3.1-70B	-.001198	-.027861	.928	9/27
Llama-3.1-405B	-.005240	-.035761	.731	8/27

Table 4: d_H minus legal uniform, paired within pool, averaged within Bemis–Murcko component, components weighted equally. No scale separates; every median is negative; the count of components favouring d_H declines with scale. Olmo-3-32B is absent because it earned no operating point.

Two further models bound feasibility rather than contributing a comparison. Qwen3-VL-2B-Instruct, the only instruction-tuned model available, reproduces the 8B legality pattern exactly under the same rule (20/20 with no observation history, 0/196 with any); its steering ladder was blocked because the longest source prompt exceeded the deployment’s attention memory. It also showed that “illegal” can bundle two things: on that tokenizer 111 of 196 top tokens were a legal label carrying a leading space, because the frozen suffix ends in a dangling space. Re-scoring the 8B cells with the same finer classification finds *no* such case: the failures there are a delimiter or an end-of-text token. The executability figures in §4 therefore do not conceal near-misses. GPT-J-6B cannot render the task at all: its 2,048-token context is shorter than the 2,638-token source prompts the construction requires.

H Audit Controls

Reference policy (42 pools)	$\Delta\tilde{V}$	p	Rank	Raw V
Masked baseline	—	—	—	.256852
EAST-inspired (frozen)	+.00285	.102	1/40	.259555
Legal uniform (untuned)	+.00247	.671	1/40	.259907
Inferential unit	$\Delta\tilde{V}$	p	Rank	n
Scaffold component	+.00139	.391	3/40	27
One pool per component	+.00069	.662	14/40	27

Table 5: Audit controls. The EAST-inspired (frozen) row reproduces the prospective, hash-frozen 1/40 result from §4; every other row is a post hoc audit computed on the same frozen cells. $\Delta\tilde{V}$ is the mean normalized value change, p a two-sided pool sign-flip test, ranks are within the same frozen bank of 39 isotropic controls plus the row’s own policy, and raw V is unnormalized oracle-continuation value. A validation-selected temperature control transfers to the reserve at rank 40/40, so flattening is not optional as a control here either (App. H).

Every quantity in Table 5 was computed from stored per-cell legal-action logits, the exact Q tables, and the frozen pool inventory with no model, tokenizer, or network calls; Table 5 itself contains no depth-balanced rows; those are reported separately in App. E from the 2,016 hosted forwards described there. Only the EAST-inspired (frozen) row is the prospective result; every other row, including legal uniform and both inferential-unit recomputations, is post hoc. The masked baseline value was reconstructed to within 10^{-9} of the stored score for all 168 cells before any control was evaluated.

The legal-uniform policy sets $\bar{\pi}_s(a) = 1/|A_s|$ so $V_s = \bar{Q}_s$; the oracle bound sets $V_s = \max_a Q_s(a)$; the temperature policy applies a softmax to the stored baseline legal logits. Uniform gains +.00247 (95% CI $[-.008294, .011368]$, $p = .671$, 28/42 pools positive) and exceeds all 39 isotropic controls. Its raw advantage over EAST-inspired despite a smaller normalized gain reflects the $1/r_s$ weighting of Eq. 2, which up-weights low- Q -range pools. The temperature grid $\{.25, .5, .75, 1.5, 2, 3, 4, 6, 8\}$ was swept on the 50-pool validation set, where $T=.25$ was best (+.013865); applied unchanged to the reserve it gives $-.003125$ and rank 40/40. A reserve-side sweep, which is selection on the outcome and reported only for completeness, peaks at $T=3$ with +.003108. We draw no positive conclusion from either number; together they show a flattening control is not optional for an entropy-derived intervention.

For the scaffold audit we compute the non-chiral RDKit Bemis–Murcko scaffold (Bemis and Mur-

cko, 1996; RDKit contributors, 2026) of every candidate SMILES, connect two pools whenever they share at least one exact nonempty scaffold, and take the transitive closure. All 1,856 candidates yield a nonempty scaffold. The 42 pools form 27 components with sizes 9, 4, 3, 2, 2 and 22 singletons. Averaging pool effects within component and weighting components equally gives +.00139 (95% CI $[-.001735, .004396]$, $p = .391$, 16/27 positive) and rank 3/40, with the same recomputation applied to every random control so the rank is like for like; keeping the lexicographically first pool per component gives +.00069 ($p = .662$) and rank 14/40. Remaining budget was recorded in the state bank and used in no selection after stage one; ranks within a stratum recompute every random control on the same pools.