SELFOCUS: MITIGATING SYSTEMATIC SELECTION BIAS IN INFERENCE-TIME SCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

Inference-time reasoning methods often rank sampled chains of thought (CoTs) using within-trace token probabilities. Final-answer frequency, however, is a many-to-one trajectory marginal. Under an idealized outcome-only, KL-regularized trajectory model, we prove that self-consistency need not recover the highest-reward answer because its ranking also depends on the total reference-policy mass Q_y of all trajectories leading to answer y. We identify a sufficient high-dimensional regime in which heterogeneous erroneous paths make a wrong answer more frequent than the uniquely rewarded answer, and demonstrate this distributional shift in a controlled multi-step experiment. We then introduce SELFOCUS, a training-free aggregator that uses cross-CoT semantic recurrence to estimate each reasoning path's deviation from the trajectory population and partially mitigate the influence of Q_y on inference-time answer selection. It clusters reasoning steps, accumulates question-local log-frequency offsets along each path, and selects the answer minimizing the ratio of mean cumulative offset to answer frequency. Across four model families and seven benchmarks, SELFOCUS improves paired majority vote in 15 of 17 model–dataset settings without token log probabilities, an answer-correctness verifier, or additional training. On larger models, measured compute also shows substantially lower GPU-hour usage than token-log-probability-based selectors.

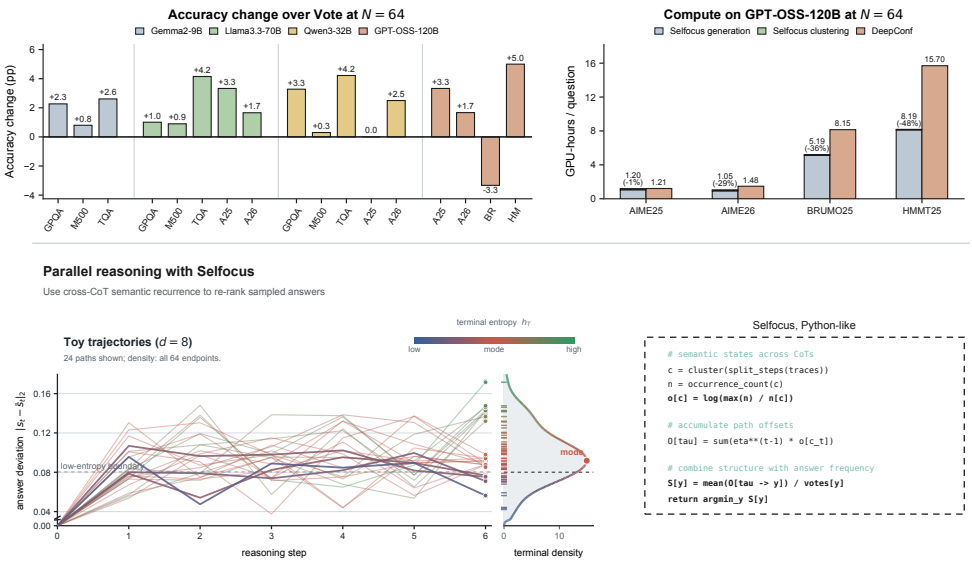


Figure 1: **Overview of SELFOCUS.** Top: accuracy changes relative to majority Vote and measured compute on GPT-OSS-120B. Bottom: recurrent semantic states across sampled CoTs identify coherent support for an answer, whereas heterogeneous deviations can inflate answer frequency. SELFOCUS aggregates this structure without additional training, token log probabilities, or an answer-correctness verifier.

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.

1 Introduction

Inference-time scaling improves language-model reasoning by sampling multiple CoTs and aggregating their final answers. We focus on improving final-answer accuracy at a fixed sampling budget without additional model training. Majority voting is a strong default, but a wrong answer can dominate by collecting unrelated failures while a correct answer has fewer, more coherent derivations. The central question is what answer frequency represents when many reasoning paths collapse to the same endpoint.

Recent work shows that generation entropy and related probability statistics contain information about reasoning quality. Existing methods aggregate token likelihoods, next-token distribution statistics, or local confidence groups into a scalar for each completed CoT. These signals are useful, but they remain local to one trajectory. A model can enter a familiar yet incorrect derivation and continue with highly predictable tokens; independently confident traces can also reach the same wrong answer for incompatible reasons. Per-trace confidence does not describe this cross-trajectory structure.

We study the problem through an idealized KL-regularized reinforcement-learning model. Starting from the Gibbs form of the optimal sequence policy, we marginalize complete trajectories by final answer. This analysis proves that self-consistency does not always recover the reward-maximizing answer: its ranking is also affected by the total reference-policy probability mass Q_y of all trajectories leading to answer y.

Our theoretical analysis proceeds in three stages. First, Gibbs marginalization establishes the dependence of answer probability on reward and reference path mass. Second, an entropy-level representation shows how Q_y can shift the answer distribution: in a sufficient high-dimensional regime, combinatorially many heterogeneous erroneous paths outweigh the correct answer's reward advantage. A controlled multi-step experiment exhibits the same shift. Third, semantic log-frequency offsets estimate trajectory deviation and partially remove the influence of Q_y from answer aggregation. Because Q_y is neither observable nor unconditionally identifiable from finite post-RL samples, we do not attempt to recover it directly. Under the stated approximation conditions, the offset instead reduces its contribution to voting, mitigating path-volume bias and improving final-answer selection during inference scaling.

This analysis motivates SelFocus, a training-free CoT-level aggregator for improving final-answer accuracy. SelFocus groups semantically recurrent steps across sampled traces, converts their relative occurrence frequencies into free-energy offsets, and accumulates these offsets along each path. Answers are ranked by mean cumulative offset normalized by answer frequency.

We evaluate SelFocus with 64 paths per question across four model families and seven reasoning benchmarks. It improves paired majority vote in 15 of 17 model–dataset cells, matches it in one, and declines in one. On larger models, it also achieves substantially greater GPU-hour savings than selectors requiring token log probabilities. Figure 1 summarizes these results; full ablations and sampling-budget scaling results appear in the supplementary material.

2 Related Work

Generation statistics for training-free selection. Generation uncertainty supports cost reduction through confidence weighting, early stopping, and path pruning (Taubenfeld et al., 2025; Zhou et al., 2025; Sharma & Chopra, 2025; Yang et al., 2026; He et al., 2026). We focus on training-free improvement of final-answer accuracy. Self-Certainty aggregates next-token distribution divergences for Best-of-N selection (Kang et al., 2025). DeepConf-Offline uses local token-group confidence to filter and reweight completed reasoning traces before answer aggregation (Fu et al., 2025). PiCSAR ranks candidates by the joint likelihood of reasoning and answer (Leang et al., 2026). These are our principal probability-based selection baselines because they share this accuracy-oriented, training-free objective. Their scores use within-trajectory probability statistics; SelFocus instead uses cross-trajectory semantic recurrence for answer re-ranking without token log probabilities. Its interpretation as partial path-volume correction is conditional on A2.

Self-consistency and path-aware aggregation. Chain-of-thought prompting exposes intermediate reasoning states (Wei et al., 2022), and self-consistency selects the most frequent answer among sam-

pled paths (Wang et al., 2023). Subsequent work extends aggregation to free-form outputs, minority-answer regimes, adaptive stopping, path pruning, and semantic weighting (Chen et al., 2023b; Yin et al., 2024; Knappe et al., 2025; Wan et al., 2025; Hong et al., 2025; Marina et al., 2026). Semantic Self-Consistency is our direct path-semantic baseline: it embeds complete CoTs and weights answer aggregation by their semantic consensus. SELFOCUS instead uses recurrent intermediate steps to re-rank answers without training a verifier, ranker, or policy. The KL-regularized analysis supplies a conditional path-volume interpretation rather than an unconditional account of the empirical gain.

KL-regularized policy theory. RLHF optimizes sequence-level reward while regularizing the learned policy toward a reference model (Ouyang et al., 2022). DPO makes the associated reward–policy log-ratio relationship explicit (Rafailov et al., 2023), and RL objectives are central to recent mathematical and CoT reasoning models (Shao et al., 2024; DeepSeek-AI, 2025). We use the Gibbs optimum of an idealized KL-regularized objective as an analytical model, not as a claim that every RL or GRPO training procedure exactly realizes this distribution.

3 ANSWER FREE ENERGY AND PATH-VOLUME BIAS

For a prompt x, let $\tau = (a_1, \dots, a_T)$ be a complete trajectory, $g(\tau) = y$ its canonical final answer, and $\mathcal{T}_y = \{\tau : g(\tau) = y\}$. We assume outcome-only reward,

$$R(\tau) = r(g(\tau)), \qquad \pi^* = \arg\max_{\pi} \mathbb{E}_{\tau\sim\pi}[R(\tau)] - \beta^{-1}\mathrm{KL}(\pi\|\rho), \tag{1}$$

where ρ is the reference trajectory distribution and $\beta > 0$.

Proposition 1 (Trajectory and answer Gibbs distributions)**.** *The optimum of Eq.* (1) *and its answer marginal satisfy*

$$\pi^*(\tau) = \frac{\rho(\tau)e^{\beta r(g(\tau))}}{Z}, \qquad p_y = \frac{Q_y e^{\beta r(y)}}{Z}, \qquad Q_y = \sum_{\tau\in\mathcal{T}_y} \rho(\tau). \tag{2}$$

Moreover, $\pi^*(\tau \mid y) = \rho(\tau \mid y)$ *and*

$$\log p_y = \beta r(y) + \log Q_y - \log Z. \tag{3}$$

Marginalization over trajectories. Summing the trajectory Gibbs policy over the fiber $\mathcal{T}_y$ gives the answer marginal in Eq. (2). Because the outcome reward is constant within each fiber, its exponential tilt cancels after conditioning. Outcome-only optimization therefore changes mass *between* answer classes while preserving relative path weights *within* each class; Q_y is exactly the inherited reference mass of all trajectories terminating at y. The complete variational derivation is given in Appendix A.

Thus the answer free energy is

$$F_y = -r(y) - \beta^{-1}\log Q_y, \qquad \arg\max_y p_y = \arg\max_y \{\beta r(y) + \log Q_y\}. \tag{4}$$

Self-consistency therefore estimates answer free energy rather than reward alone. Complete variational details are deferred to the supplementary material.

3.1 A SUFFICIENT HIGH-DIMENSIONAL FAILURE REGIME

To show that Q_y can change the maximizer, consider an effective trajectory dimension d and normalized conditional-entropy level u, with the unique correct answer $y^\star$ at $u = 0$. Suppose the number of trajectory configurations and their mean reference probability scale as $\exp\{ds(u) + o(d)\}$ and $\exp\{-de(u) + o(d)\}$. The projected level mass then has rate $\phi(u) = s(u) - e(u)$. If level $u > 0$ contains $\exp\{da(u) + o(d)\}$ canonical answers, define $\phi_{\mathrm{ans}}(u) = \phi(u) - a(u)$.

Proposition 2 (Sufficient path-volume failure)**.** *Under binary outcome reward and uniform* $o(d)$ *remainders, if* $\beta = o(d)$*, then some answer* y_d *at level* $u > 0$ *satisfies*

$$\log \frac{p_{y_d}}{p_{y^\star}} \geq d\,[\phi_{\mathrm{ans}}(u) - \phi(0)] - \beta + o(d). \tag{5}$$

Hence $\phi_{\mathrm{ans}}(u) > \phi(0)$ *is sufficient for a wrong answer eventually to exceed the uniquely rewarded answer in probability.*

Proof sketch. The total reference mass at level u scales as $e^{d\phi(u)+o(d)}$. Dividing this mass among $e^{da(u)+o(d)}$ canonical answers implies that at least one answer has reference mass $e^{d\phi_{\text{ans}}(u)+o(d)}$. Comparing it with the unique correct answer at level zero and substituting into Eq. (2) gives Eq. (5). Appendix A gives the complete level-concentration and answer-level arguments.

This is one sufficient failure regime, not an unconditional claim about real language models. It formalizes the mechanism of interest: the number and reference mass of heterogeneous erroneous trajectories can grow faster than the reward advantage. If Q_y were available, the strict correction would be

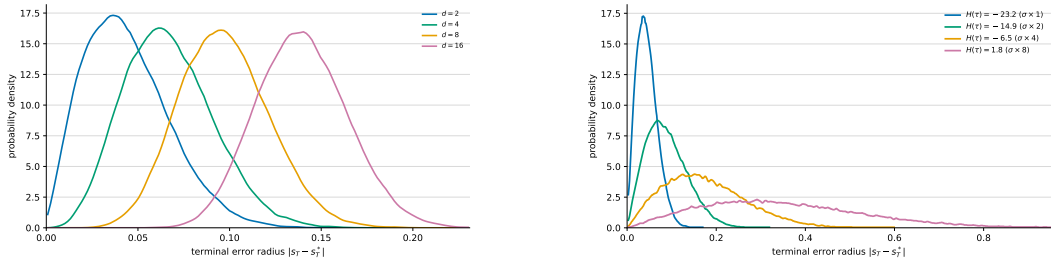
$$U_y = \log p_y - \log Q_y = \beta r(y) - \log Z. \tag{6}$$

In sample-only inference, however, Q_y is unobserved, motivating an estimator based on semantic trajectory recurrence.

3.2 Controlled demonstration of the shift

We test the geometric mechanism in a continuous six-step point-finding problem. For $d \in \{2, 4, 8, 16\}$, the state begins at the origin, the target is e_1, and actions are Gaussian increments with a two-layer MLP mean. A separate policy in each dimension is trained by REINFORCE for 2,500 updates with batch size 1,024 and terminal reward $-\|s_T - e_1\|^2/[2(0.35)^2]$. We then draw 60,000 trajectories per condition. All dimension-sweep policies use per-coordinate evaluation standard deviation 0.035; their deterministic mean trajectories terminate within 0.004–0.017 of the target, and the trajectory differential-entropy rate remains between -1.934 and -1.932.

Figure 2 reports both controlled interventions. At fixed local stochasticity, the mean terminal radius rises from 0.044 at $d = 2$ to 0.067, 0.097, and 0.139 at $d = 4, 8, 16$. At fixed $d = 2$, we freeze the learned mean policy and multiply the sampling standard deviation by $1, 2, 4, 8$. The estimated trajectory entropy changes from -23.2 to -14.9, -6.5, 1.8, while the mean terminal radius rises from 0.044 to 0.089, 0.176, 0.353. Increasing either effective dimension or local dispersion therefore shifts aggregate trajectory mass away from a reward-aligned mean path. The experiment isolates this geometric effect; it does not itself instantiate the discrete-answer Gibbs model.



(a) Fixed per-coordinate dispersion; dimension varies.

(b) Fixed dimension; sampling dispersion varies.

Figure 2: **Controlled trajectory-distribution shift.** The learned mean trajectory remains near the target, but the sampled terminal distribution moves outward as either effective trajectory dimension or local stochasticity increases.

To examine A2 in a setting where reference mass is observable, we fix the semantic-state and answer maps before estimating either Q_y or G_y. Semantic states discretize each step's deviation from the deterministic mean trajectory by normalized radius and fixed direction prototypes; final answers comprise one correct ball and 64 wrong regions formed by four radial shells and 16 directional sectors. Independent million-trajectory splits estimate Q_y and G_y. Calibration-to-heldout sub-Gaussian scale ratios remain near one for $d = 4$ and $d = 8$, while the absolute residual scale grows sharply at $d = 16$. The generic sufficient ranking condition derived below holds in every partition for four of the six supported dimension–discount settings, and the exact population corrected-score margin is positive in every partition for all six. Appendix D.2 gives the complete importance-weighted construction and further compares Vote, population SELFOCUS, and the oracle p_y/Q_y. This is a controlled diagnostic of A2, not evidence that the benchmark gains identify Q_y.

4 SELFOCUS: SEMANTIC TRAJECTORY AGGREGATION

The strict score in Eq. (6) requires reference-policy answer mass. SELFOCUS instead estimates relative semantic deviation from post-RL samples. For each question, it parses every CoT into ordered reasoning steps and uses a fixed two-stage O/C procedure to construct a sample-dependent semantic partition. Stage 1 clusters semantically overlapping occurrences only at the same step index across trajectories. Stage 2 then compares the resulting clusters across step indices and merges semantic overlap or containment. Embedding retrieval proposes candidate pairs, while a bidirectional NLI classifier makes the merge decision without judging correctness. The same procedure is used for every model and benchmark.

Let n_c be the number of raw occurrences of cluster c, counted before path compression, and let $c^* = \arg\max_c n_c$. The empirical step offset is

$$\widehat{o}(c) = \log \frac{n_{c^*}}{n_c}. \tag{7}$$

Let $\mu(c)$ be the population occurrence probability and $\mu_* = \max_c \mu(c)$. With $F_c^{\mathrm{occ}} = -\beta^{-1} \log \mu(c)$ up to a common additive constant and $F_*^{\mathrm{occ}} = \min_c F_c^{\mathrm{occ}}$, the population offset obeys

$$o(c) = \log \frac{\mu_*}{\mu(c)} = \beta \left(F_c^{\mathrm{occ}} - F_*^{\mathrm{occ}} \right). \tag{8}$$

Since $\pi(\tau \mid y) = \rho(\tau \mid y)$, the answer-conditioned population offset mean is unchanged by the outcome-only reward tilt.

Adjacent repeated cluster IDs are removed after occurrence counting. For the compressed path $(c_1, \ldots, c_{\widetilde{L}})$, define

$$\widehat{O}_\eta(\tau) = \sum_{t=1}^{\widetilde{L}} \eta^{t-1} \widehat{o}(c_t), \qquad \overline{O}_y = \frac{1}{n_y} \sum_{\tau : g(\tau) = y} \widehat{O}_\eta(\tau), \qquad S(y) = \frac{\overline{O}_y}{n_y}. \tag{9}$$

The selected answer minimizes $S(y)$ over all distinct normalized answers sampled for the question. This rule is identical across datasets.

Label-free depth discount. One run-level coefficient is estimated from a fixed-seed unlabeled calibration subset containing $\lceil 0.3|Q| \rceil$ questions for every benchmark. Let R be the question-balanced rate at which post-initial semantic states receive repeated support from another path, and let D be the mean increase in offset from the first to the second half of a path. We use

$$\eta = \Pi_{\{0.1, \ldots, 1.0\}} \left(\mathrm{clip}_{[0.1, 1]} [R e^{-D}] \right), \tag{10}$$

with an exact grid midpoint mapped to the smaller value. The coefficient uses no answer or correctness label. It is a structural aggregation heuristic rather than a consequence of Eq. (3).

Occurrence consistency and the residual link to path volume. Let $q_y := \overline{O}_y$ be the implemented answer-level offset and define $G_y := \mathbb{E}_{\pi(\cdot|y)} [O^\circ_\eta(\tau)]$ under a sample-independent reference semantic partition. Under bounded occurrence counts and path lengths and positive population occupancy, if the statistic discrepancy induced by the actual O/C partition is at most κ_N with high probability, then

$$|q_y - G_y| \le \varepsilon_{N,y} + \kappa_N, \qquad \varepsilon_{N,y} = O_p(N^{-1/2}). \tag{11}$$

The sampling rate holds for answers of fixed positive probability; consistency of the implemented statistic additionally requires $\kappa_N \to 0$.

Relating G_y to reference path mass requires the reward-conditioned residual assumption A2:

$$\log \frac{G_y}{Q_y} = h_x(r(y)) + \xi_y, \qquad |\xi_y| \le \epsilon_x. \tag{12}$$

Together with the answer Gibbs distribution, this gives

$$\log \frac{p_y}{G_y} = \Phi_x(r(y)) - \log Z - \xi_y, \qquad \Phi_x(u) = \beta u - h_x(u). \tag{13}$$

The correction thus leaves a reward-dependent term and an unresolved structural residual, rather than eliminating path-mass effects completely.

For the empirical score $\widehat{s}_y = \log[(n_y/N)/q_y]$, when estimation errors are smaller than their corresponding population quantities,

$$\left|\widehat{s}_y - [\Phi_x(r(y)) - \log Z - \xi_y]\right| \le D_{N,y}, \qquad D_{N,y} = O_p(N^{-1/2} + \kappa_N). \tag{14}$$

The rate assumes fixed positive supports. For the unique rewarded correct answer $y^\star$, let $m_\Phi = \min_{y \neq y^\star}\{\Phi_x(r(y^\star)) - \Phi_x(r(y))\}$. On the event where these bounds hold, if $y^\star$ is sampled and

$$m_\Phi > 2\epsilon_x + D_{N,y^\star} + \max_{y \neq y^\star} D_{N,y}, \tag{15}$$

then $y^\star$ uniquely maximizes $\widehat{s}_y$, equivalently minimizing q_y/n_y. This is a sufficient, not unconditional, ranking guarantee. More samples reduce sampling error but do not automatically remove clustering bias or structural residuals. Appendix B.3 gives the complete bounds, proofs, and finite-grid extension for η; Appendix D.2 provides controlled residual diagnostics.

Algorithm 1 SELFOCUS inference

Require: Questions $\mathcal{Q}$, sampled CoTs $\mathcal{T}_q$, encoder E, NLI classifier V, seed ξ
1: **for** each question q **do**
2: Parse answers and reasoning steps; construct fixed O/C semantic clusters with E, V.
3: Count raw cluster occurrences, compute $\widehat{o}_q(c)$, then adjacent-deduplicate each path.
4: **end for**
5: Estimate one run-level η from the fixed-seed unlabeled subset defined above using Eq. (10).
6: **for** each question q **do**
7: Compute path offsets and answer scores $S_q(y) = \overline{O}_{q,y}/n_{q,y}$.
8: Select the minimum-score answer set over all distinct normalized answers.
9: **end for**
10: **return** selected answers.

5 EXPERIMENTS

Setup. We evaluate Gemma2-9B-IT, Llama3.3-70B-Instruct, Qwen3-32B, and GPT-OSS-120B with 64 sampled paths per question. Benchmarks comprise the 198-question GPQA Diamond split (Rein et al., 2024), MATH500 (Hendrycks et al., 2021), 747 text-only TheoremQA questions (Chen et al., 2023a), AIME25, AIME26, BRUMO25, and HMMT25. Every model–dataset experiment is reported over four independent runs. All experiments are run on a server equipped with eight NVIDIA A100 GPUs. The main results use the η values returned by the unsupervised, label-free estimator in Eq. (10); fixed $\eta = 1$, component ablations, and complete sampling-budget curves are evaluated in the supplementary material. Complete prompts, decoding configurations, semantic path construction, and answer processing are also reported there.

Baselines. We compare SELFOCUS with majority Vote, DeepConf-Offline (Fu et al., 2025), Self-Certainty (Kang et al., 2025), PiCSAR (Leang et al., 2026), and Semantic Self-Consistency (SSC) (Knappe et al., 2025). We implement SSC with the 32k-context GTE-Qwen2-1.5B-Instruct encoder (Li et al., 2023). Complete baseline implementations and the candidate-pool protocol are provided in the supplementary material.

5.1 MAIN RESULTS

Trajectory-level structural signal. Before comparing answer selectors, we test whether occurrence-space structure is associated with trajectory correctness. Let $C^{\text{raw}}(\tau) = (c_1, \ldots, c_L)$ be the uncompressed cluster sequence and $\widetilde{C}(\tau) = (\widetilde{c}_1, \ldots, \widetilde{c}_{\widetilde{L}})$ its adjacent-deduplicated path. We define

$$\text{Offset}_{\text{mean}}(\tau) = \frac{1}{\widetilde{L}} \sum_{t=1}^{\widetilde{L}} \widehat{o}(\widetilde{c}_t), \qquad \text{ASC}(\tau) = \frac{1}{L} \sum_{t=1}^{L} \frac{n_{c_t}}{n_{c^*}}. \tag{16}$$

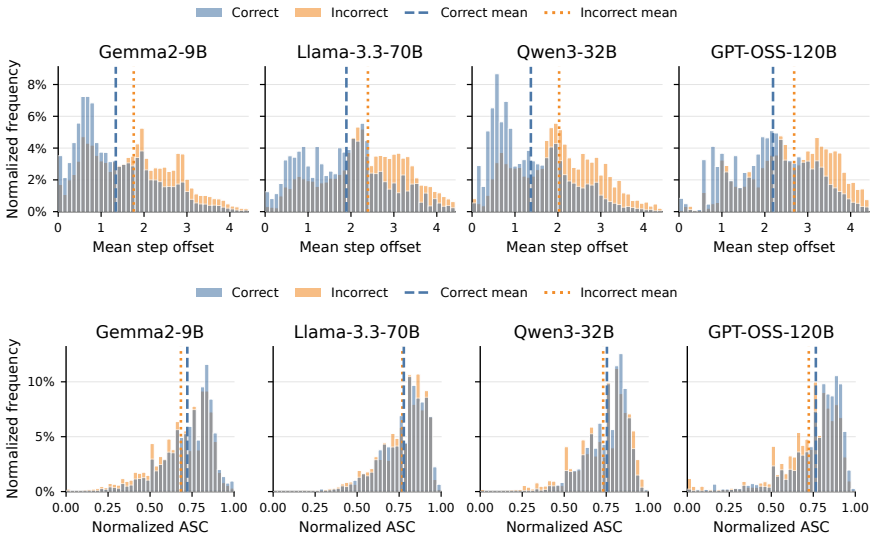


Figure 3: **Trajectory-level structural signals by model.** Distributions are normalized separately within each correctness class. Top: correct CoTs have lower mean step offset. Bottom: correct CoTs have higher normalized ASC, defined as the mean visited-cluster size divided by the modal cluster size of the same question-run. Dashed and dotted vertical lines mark the correct and incorrect class means, respectively.

The first statistic is the mean step offset at $\eta = 1$; the second is normalized answer semantic concentration over raw step occurrences. Figure 3 shows that, across all four model families, correct CoTs have lower mean step offset and higher normalized ASC than incorrect CoTs. The mean offset gap ranges from 0.42 to 0.66, while the normalized-ASC gap ranges from 0.009 to 0.042. Although the distributions overlap, their consistent directional shift shows that cross-trajectory semantic recurrence contains an outcome-related signal before answer-level aggregation.

Answer-selection accuracy. Across the 17 model–dataset cells in Table 1, SELFOCUS improves paired Vote in 15, matches it in one, and declines in one. Against SSC, which also aggregates cross-trajectory semantic information, SELFOCUS is higher in 15 cells, equal in one, and lower in one, with a macro-average advantage of 1.68 percentage points. The improvement is not confined to a particular model scale or initial accuracy level. On the three benchmarks shared by Gemma, Llama, and Qwen, SELFOCUS improves Vote in all nine comparisons and achieves the highest reported mean among the evaluated selectors. The gains are modest but consistent on MATH500, larger on GPQA, and strongest on TheoremQA, where they range from 2.61 to 4.22 percentage points. The same pattern extends to the high-difficulty suite: GPT-OSS gains 5.00 points on HMMT25 and improves on both AIME sets, although it declines on BRUMO25. Overall, the results show that cross-trajectory semantic recurrence is useful for answer re-ranking across architectures and benchmarks. They do not by themselves identify path-volume correction as the cause of the gain.

Paired robustness. Figure 4 distinguishes targeted corrections from simply exchanging errors through paired question-level transitions. Correction rates substantially exceed regression rates across nearly all model–dataset settings, while SELFOCUS preserves 98.9% of the decisions that Vote already answers correctly. The gains therefore arise primarily from correcting erroneous majorities while retaining successful decisions, supporting the robustness of the aggregation rule. Complete transition rates appear in Appendix E.2.

Table 1: **Accuracy on reasoning benchmarks.** Results are reported as mean accuracy (%) ± sample standard deviation over four independent runs (percentage points). The highest available mean in each row is bolded. SSC denotes Semantic Self-Consistency. GPT-OSS-120B is evaluated separately on the high-difficulty suite.

	Baselines					Ours
Dataset	Vote	DeepConf	Self-Certainty	PiCSAR	SSC	SELFOCUS
Gemma2-9B						
GPQA	37.12 ±1.46	38.38 ±1.43	27.40 ±1.19	33.96 ±2.75	37.63 ±1.46	**39.39** ±1.75
MATH500	54.70 ±0.81	–	43.95 ±0.10	47.15 ±0.19	55.40 ±0.92	**55.50** ±1.50
TheoremQA	40.36 ±0.70	–	41.83 ±1.47	37.22 ±1.45	41.16 ±0.85	**42.97** ±0.78
Llama3.3-70B						
GPQA	47.73 ±1.46	48.48 ±1.43	38.89 ±0.92	46.46 ±3.67	48.24 ±0.87	**48.74** ±0.87
MATH500	75.50 ±0.12	–	61.70 ±0.20	64.05 ±1.23	75.70 ±0.12	**76.40** ±0.00
TheoremQA	65.06 ±0.78	–	67.54 ±0.08	64.26 ±1.70	66.26 ±0.39	**69.21** ±0.31
AIME25	10.00 ±0.00	3.33 ±0.00	6.67 ±3.85	6.67 ±3.85	10.00 ±0.00	**13.33** ±0.00
AIME26	16.67 ±0.00	10.00 ±2.72	13.33 ±4.71	10.00 ±4.71	16.67 ±0.00	**18.33** ±1.92
Qwen3-32B						
GPQA	59.09 ±0.00	61.11 ±1.43	43.43 ±1.01	44.44 ±3.22	60.10 ±1.17	**62.37** ±1.46
MATH500	86.20 ±0.00	–	75.70 ±0.12	75.95 ±0.19	86.40 ±0.23	**86.50** ±0.12
TheoremQA	77.51 ±0.15	–	80.19 ±0.93	73.56 ±0.70	79.12 ±0.31	**81.73** ±0.23
AIME25	80.00 ±6.09	**80.83** ±3.19	80.00 ±6.09	70.00 ±6.67	80.00 ±0.00	80.00 ±6.09
AIME26	85.00 ±1.92	**90.00** ±2.72	**90.00** ±4.71	73.33 ±6.67	85.00 ±3.85	87.50 ±1.67

GPT-OSS-120B High-difficulty suite

Dataset	Vote	DeepConf	SSC	SELFOCUS
AIME25	81.67 ±1.92	**96.67** ±0.00	81.67 ±1.92	85.00 ±5.77
AIME26	93.33 ±0.00	93.33 ±2.72	93.33 ±0.00	**95.00** ±1.92
BRUMO25	**78.33** ±5.77	76.67 ±2.72	**78.33** ±5.77	75.00 ±5.77
HMMT25	61.67 ±1.92	63.33 ±3.85	60.00 ±0.00	**66.67** ±3.85

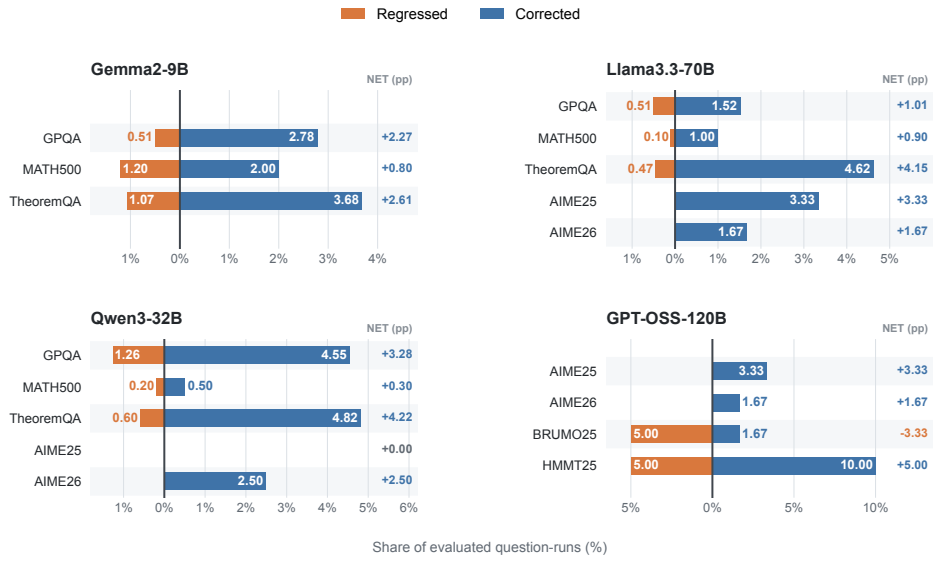


Figure 4: **Paired answer transitions from Vote to SELFOCUS.** Rightward bars show the percentage of evaluated question-runs corrected by SELFOCUS; leftward bars show the percentage regressed from a correct Vote answer. Net values are accuracy changes in percentage points. Each model panel uses its own horizontal scale.

5.2 Compute usage

Unlike token-probability selectors, Selfocus requires neither per-token output log probabilities nor an additional teacher-forced scoring pass. Its extra cost after generation is semantic clustering. Figure 5 compares per-question GPU-hours at $N = 64$ and separates these two costs.

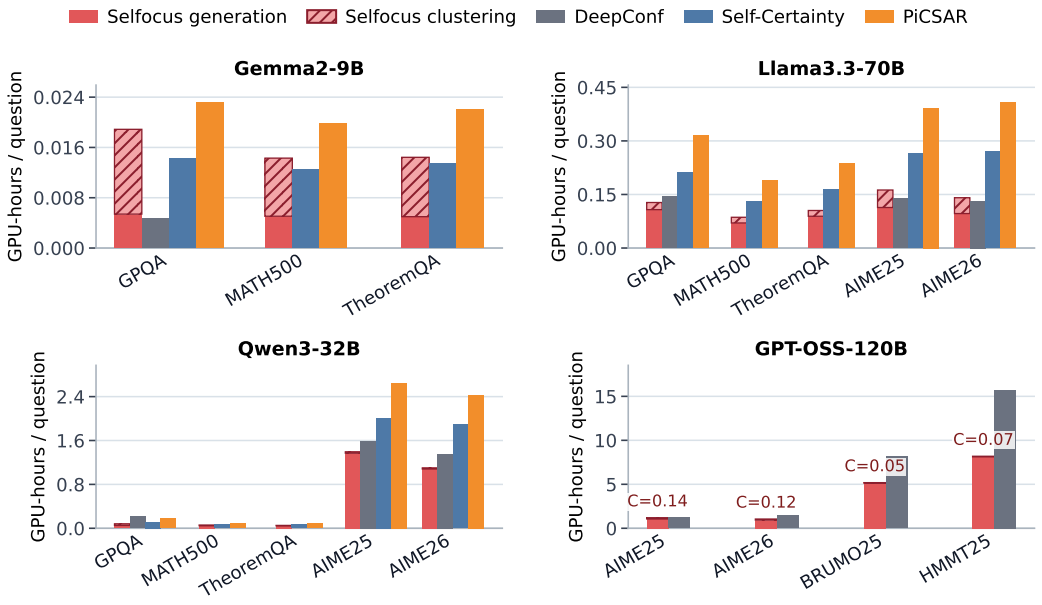


Figure 5: **Compute usage by model and dataset.** GPU-hours per question at $N = 64$, measured on the same eight-A100 server. Each Selfocus bar separates generation and semantic clustering. Panels use independent linear vertical scales; C labels clustering cost where its segment is small.

Clustering dominates the relatively inexpensive Gemma2-9B pipeline, but becomes a small share as generation grows more costly. For Qwen3-32B AIME26, clustering is 0.0180 of 1.0992 GPU-hours per question, and the total is lower than DeepConf (1.3556), Self-Certainty (1.8943), and PiCSAR (2.4330). On GPT-OSS-120B, Selfocus uses 5.19 versus 8.15 GPU-hours on BRUMO25 and 8.19 versus 15.70 on HMMT25, reducing cost relative to DeepConf by approximately 36% and 48%. The measured savings thus become substantial in the larger-model, generation-intensive settings, despite the added clustering stage.

6 Conclusion

We study path-mass bias in majority-vote inference scaling. Marginalizing an idealized KL-regularized trajectory policy by final answer shows that answer frequency depends on both reward and the total reference-policy path mass Q_y. This motivates Selfocus, a training-free aggregator that uses cross-CoT semantic recurrence to compute path offsets and partially mitigate the influence of Q_y on answer ranking under the stated conditions.

Across four model families, correct trajectories exhibit lower mean step offsets and higher semantic concentration. Selfocus improves Vote in 15 of 17 settings, mainly by correcting erroneous majorities while preserving successful decisions. Ablations favor offset-based aggregation over pure ASC and ASC-weighted Vote, show the importance of answer-frequency normalization, and find similar performance for Mean and Full, suggesting that cumulative path length is not the main source of the gain. Finally, avoiding token log probabilities and teacher-forced scoring yields substantial GPU-hour savings in larger-model settings, where clustering accounts for a small share of total cost.

AI USE STATEMENT

Generative AI tools were used to assist with formulating and checking mathematical claims and proofs, and implementing methods and experimental code. The authors reviewed all AI-assisted work and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The main paper states the assumptions and decision rule used by SELFOCUS. The supplementary material contains complete proofs, expanded pseudocode, the unified prompts and experimental protocol, additional trajectory-level analyses, all model–dataset ablations, and sampling-budget results.

REFERENCES

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in LLM reasoning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 107150–107180. Curran Associates, Inc., 2025. doi: 10.52202/085713-3573. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/9a14cf2ee36cf2524f54d8f196a6734c-Paper-Conference.pdf.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. TheoremQA: A theorem-driven question answering dataset. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7889–7901, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.489. URL https://aclanthology.org/2023.emnlp-main.489/.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language model generation, 2023b. URL https://arxiv.org/abs/2311.17311.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence, 2025. URL https://arxiv.org/abs/2508.15260.

Jiashu He, Meizhu Liu, Olaitan P Olaleye, Amit Agarwal, M. Avendi, Yassi Abbasi, Matthew Rowe, Hitesh Laxmichand Patel, Paul Li, Tao Sheng, Sujith Ravi, and Dan Roth. Think twice before you write – an entropy-based decoding strategy to enhance LLM reasoning, 2026. URL https://arxiv.org/abs/2604.00018.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf.

Colin Hong, Xu Guo, Anand Chaanan Singh, Esha Choukse, and Dmitrii Ustiugov. Slim-SC: Thought pruning for efficient scaling with self-consistency. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34500–34517, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1750. URL https://aclanthology.org/2025.emnlp-main.1750/.

Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-N selection for large language models via self-certainty. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 19720–19745. Curran Associates, Inc., 2025. doi: 10.52202/085713-0663. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/1c7eff166a8e345f664f0faa8f4e4d2e-Paper-Conference.pdf.

Tim Knappe, Ryan Li, Ayush Chauhan, Kaylee Chhua, Kevin Zhu, and Sean O'Brien. Semantic self-consistency: Enhancing language model reasoning via semantic weighting, 2025. URL https://arxiv.org/abs/2410.07839.

Joshua Ong Jun Leang, Zheng Zhao, Aryo Pradipta Gema, Sohee Yang, Wai-Chung Kwan, Xuanli He, Wenda Li, Pasquale Minervini, Eleonora Giunchiglia, and Shay B Cohen. PiCSAR: Probabilistic confidence selection and ranking for reasoning chains. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 31511–31544, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1577. URL https://aclanthology.org/2026.findings-acl.1577/.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023. URL https://arxiv.org/abs/2308.03281.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf.

Maria Marina, Daniil Moskovskiy, Sergey Pletenev, Mikhail Salnikov, Alexander Panchenko, and Viktor Moskvoretskii. Boosting self-consistency with ranking. In Santosh T.Y.S.S., Juan Diego Rodriguez, and Ona de Gibert (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 1017–1032, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-393-7. doi: 10.18653/v1/2026.acl-srw.89. URL https://aclanthology.org/2026.acl-srw.89/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

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. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of

mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Aman Sharma and Paras Chopra. Think just enough: Sequence-level entropy as a confidence signal for LLM reasoning, 2025. URL https://arxiv.org/abs/2510.08146.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence improves self-consistency in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20090–20111, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1030. URL https://aclanthology.org/2025.findings-acl.1030/.

Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. Reasoning aware self-consistency: Leveraging reasoning paths for efficient LLM sampling. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 3613–3635, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.184. URL https://aclanthology.org/2025.naacl-long.184/.

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 *ICLR 2023*, 2023. URL https://arxiv.org/abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2550–2575, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.167. URL https://aclanthology.org/2023.findings-emnlp.167/.

Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen, Zheng Lin, and Weiping Wang. Dynamic early exit in reasoning models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 88170–88210, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/8df90a1440ce782d1f5607b7a38f2531-Paper-Conference.pdf.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Zhiyuan Zeng, Xiaonan Li, Tianxiang Sun, Cheng Chang, Qinyuan Cheng, Ding Wang, Xiaofeng Mou, Xipeng Qiu, and Xuanjing Huang. Aggregation of reasoning: A hierarchical framework for enhancing answer selection in large language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 609–625, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.53/.

Kongcheng Zhang, Qi Yao, Shunyu Liu, Yingjie Wang, Baisheng Lai, Jieping Ye, Mingli Song, and Dacheng Tao. Consistent paths lead to truth: Self-rewarding reinforcement learning for LLM reasoning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 59849–59887. Curran Associates, Inc., 2025. doi: 10.52202/085713-2002. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/565f995643da6329cec701f26f8579f5-Paper-Conference.pdf.

Zhi Zhou, Yuhao Tan, Zenan Li, Yuan Yao, Lan-Zhe Guo, Yu-Feng Li, and Xiaoxing Ma. A theoretical study on bridging internal probability and self-consistency for LLM reasoning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 87380–87413. Curran Associates, Inc., 2025. doi: 10.52202/085713-2922. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/7e9afa9a02857bce4515247842471444-Paper-Conference.pdf.

A Complete Theoretical Derivations

For a prompt x, autoregressive decoding induces a distribution over complete reasoning trajectories,

$$\pi(\tau \mid x) = \prod_{t=1}^{T} p(a_t \mid x, a_{<t}), \qquad \tau = (a_1, \ldots, a_T). \tag{17}$$

Let $g : \mathcal{T} \to \mathcal{Y}$ map a trajectory to its final answer and let $\mathcal{T}_y = \{\tau : g(\tau) = y\}$. Self-consistency estimates the answer marginal

$$p_y = \Pr_{\tau \sim \pi}[g(\tau) = y] = \sum_{\tau \in \mathcal{T}_y} \pi(\tau) \tag{18}$$

by n_y/N and returns the largest estimated marginal. We characterize this marginal under a KL-regularized trajectory model, identify the post-RL occurrence functional estimated from samples, and state conditions under which it partially captures the path-volume term.

A.1 KL-Regularized Trajectory Distribution

Assume that the terminal reward assigned to a complete trajectory depends only on its canonical final answer:

$$R(\tau) = r(g(\tau)). \tag{19}$$

Thus all trajectories in $\mathcal{T}_y$ receive the same reward $r(y)$.

Consider the KL-regularized objective

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau \sim \pi}[r(g(\tau))] - \frac{1}{\beta}\mathrm{KL}(\pi \| \rho), \tag{20}$$

where ρ is a reference trajectory distribution and $\beta > 0$ is an inverse temperature.

Proposition 3 (Gibbs form of the KL-regularized optimum)**.** *The optimizer of Eq. 20 satisfies*

$$\pi^*(\tau) = \frac{\rho(\tau)\exp(\beta r(g(\tau)))}{Z}, \quad Z = \sum_{\tau \in \mathcal{T}} \rho(\tau)\exp(\beta r(g(\tau))). \tag{21}$$

Proof. Stationarity of the normalized Lagrangian gives $r(g(\tau)) - \beta^{-1}(\log(\pi(\tau)/\rho(\tau)) + 1) + \lambda = 0$. Exponentiating and normalizing yields Eq. 21. □

The subsequent analysis assumes the optimum in Eq. 21.

A.2 Answer Free Energy and Path-Volume Bias

Under Eq. 21, the answer marginal is

$$p_y = \sum_{\tau \in \mathcal{T}_y} \pi(\tau) = \frac{\exp(\beta r(y))}{Z} \sum_{\tau \in \mathcal{T}_y} \rho(\tau) = \frac{Q_y \exp(\beta r(y))}{Z} = \frac{Z_y}{Z}, \tag{22}$$

where

$$Q_y = \sum_{\tau \in \mathcal{T}_y} \rho(\tau), \qquad Z_y = Q_y \exp(\beta r(y))$$

are the reference trajectory mass and restricted partition function for answer y.

Define $\rho_y(\tau) = \rho(\tau)/Q_y$ and $\pi_y(\tau) = \pi(\tau \mid g(\tau) = y)$. Then:

Proposition 4 (Outcome-reward answer decomposition)**.** *For any answer y with $p_y > 0$ and $Q_y > 0$,*

$$\pi_y(\tau) = \rho_y(\tau), \qquad \log p_y = \beta r(y) + \log Q_y - \log Z. \tag{23}$$

Proof. For every $\tau \in \mathcal{T}_y$, the reward multiplier is the same constant $\exp(\beta r(y))$. It therefore cancels under conditioning:

$$\pi_y(\tau) = \frac{\rho(\tau)\exp(\beta r(y))}{Q_y \exp(\beta r(y))} = \rho_y(\tau).$$

Taking the logarithm of Eq. 22 gives the second identity. □

Define the answer free energy

$$F_y = -\frac{1}{\beta}\log Z_y = -r(y) - \frac{1}{\beta}\log Q_y.$$

Then $p_y = \exp(-\beta F_y)/\sum_{y'}\exp(-\beta F_{y'})$: the answer distribution is Gibbs with respect to answer free energy, not reward alone. The exact path-volume bias is

$$B_y^\rho = \log Q_y.$$

Consequently,

$$\arg\max_y p_y = \arg\max_y \{\beta r(y) + \log Q_y\},$$

so self-consistency need not recover the reward-maximizing answer.

A.3 A Sufficient Entropy-Projection Failure Regime

We now give a sufficient regime in which Q_y changes the maximizer of the answer distribution. The regime assumes a sparse correct basin and combinatorially many heterogeneous deviations mapped to a smaller set of canonical wrong answers.

For each answer y, define its conditional trajectory entropy

$$h_y = H(\rho_y) = -\mathbb{E}_{\tau\sim\rho_y}[\log \rho_y(\tau)]. \tag{24}$$

Proposition 4 gives $H(\pi_y) = H(\rho_y) = h_y$.

Assumption 1 (Entropy-ordered hard-task regime)**.** *The answer set can be partitioned into ordered entropy levels $\{\mathcal{Y}_k\}_{k=0}^K$, where $\mathcal{Y}_k = \{y : h_y \in I_k\}$ for intervals $I_0 < I_1 < \cdots < I_K$. The unique correct answer $y^\star$ forms the lowest level, $\mathcal{Y}_0 = \{y^\star\}$, while higher levels collect answers supported by increasingly heterogeneous combinations of reasoning deviations.*

This is a structural assumption on the task and the reference distribution.

Define the reference mass of entropy level k by

$$Q_k = \sum_{y\in\mathcal{Y}_k} Q_y. \tag{25}$$

Let K_H denote the entropy level of an answer sampled from p_y. Pushing Eq. 22 through the level map gives the exact distribution

$$\Pr(K_H = k) = \frac{\sum_{y\in\mathcal{Y}_k} Q_y \exp(\beta r(y))}{\sum_j \sum_{y\in\mathcal{Y}_j} Q_y \exp(\beta r(y))}. \tag{26}$$

For binary outcome reward, $r(y) = \mathbb{I}\{y = y^\star\}$, this reduces to

$$\Pr(K_H = k) = \frac{Q_k \exp(\beta \mathbb{I}\{k=0\})}{Q_0 e^\beta + \sum_{j=1}^K Q_j}. \tag{27}$$

The correct level receives log-weight β, while Q_k carries the projected reference mass of level k.

To characterize one high-dimensional realization of this regime, let d denote the number of effective reasoning degrees of freedom and let $u = k/d$ be a normalized entropy level, with the number of levels allowed to depend on d. Let $\mathcal{U}_d \subset [0, u_{\max}]$ be the finite set of supported normalized levels, where $u_{\max} < \infty$ is fixed and $0 \in \mathcal{U}_d$. Suppose that the effective number of trajectory configurations and their mean reference probability satisfy, uniformly over $u \in \mathcal{U}_d$,

$$N_d(u) = \exp\{ds(u) + o(d)\}, \qquad \overline{\rho}_d(u) = \exp\{-de(u) + o(d)\}. \tag{28}$$

The projected reference mass then obeys

$$Q_d(u) = N_d(u)\overline{\rho}_d(u) = \exp\{d\phi(u) + \zeta_d(u)\}, \qquad \phi(u) = s(u) - e(u), \tag{29}$$

where $s(u)$ is the combinatorial growth rate, $e(u)$ is the mean single-trajectory surprisal rate, and $\|\zeta_d\|_\infty = \sup_{u\in\mathcal{U}_d} |\zeta_d(u)| = o(d)$. In this asymptotic regime, $\beta = \beta_d > 0$ may depend on d.

Proposition 5 (Entropy-induced level shift)**.** *Under binary outcome reward, the normalized entropy level $U = K_H/d$ has distribution*

$$\Pr_d(U = u) = \frac{\exp\{d\phi(u) + \beta\mathbb{I}\{u = 0\} + \zeta_d(u)\}}{\sum_{v\in\mathcal{U}_d} \exp\{d\phi(v) + \beta\mathbb{I}\{v = 0\} + \zeta_d(v)\}}. \tag{30}$$

Assume $\log|\mathcal{U}_d| = o(d)$, $\beta = o(d)$, and that ϕ is continuous on $[0, u_{\max}]$ with a unique maximizer $u^\dagger > 0$. If there exists $v_d \in \mathcal{U}_d$ with $v_d \to u^\dagger$, then, for every $\varepsilon > 0$,

$$\Pr_d\Big(|U - u^\dagger| \geq \varepsilon\Big) \to 0 \qquad (d \to \infty). \tag{31}$$

In particular, $\Pr_d(U = 0) \to 0$, although $u = 0$ is the unique rewarded level.

Proof. Substituting Eq. 29 into Eq. 27 gives Eq. 30. Fix $\varepsilon > 0$ and let $E_\varepsilon = \{u \in [0, u_{\max}] : |u - u^\dagger| \geq \varepsilon\}$. If this set is empty, the conclusion is immediate. Otherwise, continuity, compactness, and uniqueness imply

$$\Delta_\varepsilon = \phi(u^\dagger) - \sup_{u\in E_\varepsilon} \phi(u) > 0.$$

The denominator contains the contribution of v_d, with $\phi(v_d) = \phi(u^\dagger) + o(1)$. Bounding the numerator over at most $|\mathcal{U}_d|$ levels yields

$$\begin{aligned}\Pr_d(U \in E_\varepsilon) &\leq |\mathcal{U}_d| \exp\left\{d\left[\sup_{u\in E_\varepsilon} \phi(u) - \phi(v_d)\right] + \beta + 2\|\zeta_d\|_\infty\right\} \\ &= \exp\{-d\Delta_\varepsilon + o(d)\} \to 0.\end{aligned}$$

Taking $0 < \varepsilon < u^\dagger$ also gives $\Pr_d(U = 0) \to 0$. □

To obtain an answer-level condition, consider supported levels $u_d \to u > 0$ containing $A_d(u_d) = \exp\{da(u) + o(d)\}$ canonical answers. Let y_d be an answer with maximal reference mass within level u_d. Then $Q_{y_d} \geq Q_d(u_d)/A_d(u_d)$, whose lower exponential rate is

$$\phi_{\text{ans}}(u) = s(u) - e(u) - a(u). \tag{32}$$

Because $\mathcal{Y}_0 = \{y^\star\}$, continuity of ϕ and the uniform remainder bound give

$$\log\frac{p_{y_d}}{p_{y^\star}} \geq d[\phi_{\text{ans}}(u) - \phi(0)] - \beta + o(d). \tag{33}$$

If $\phi_{\text{ans}}(u) > \phi(0)$ and $\beta = o(d)$, this lower bound tends to $+\infty$, so at least one wrong answer eventually has greater Gibbs mass than $y^\star$. This is a sufficient, not necessary, condition for self-consistency failure induced by Q_y.

A.4 Strict Reward-Aligned Correction

If the reference policy is available, Eq. 23 gives the strict reward-aligned score

$$U_y = \log p_y - \log Q_y = \beta r(y) - \log Z. \tag{34}$$

In our setting only samples from π are available. Reusing them to estimate Q_y directly is invalid because they carry post-RL mass p_y, not reference mass. Equation 34 identifies the ideal correction. The analysis below instead determines what the observable semantic statistic estimates and isolates the residual assumptions required to connect it to Q_y.

B Full Derivation of Selfocus

B.1 Estimating Semantic Occurrence and Relating It to Path Volume

The strict correction in Eq. 34 requires the reference-policy answer mass Q_y, which integrates the mass of all trajectories terminating at answer y. This quantity is unavailable when inference provides only a finite collection of post-RL CoT samples, and it cannot be recovered by answer frequency alone. We therefore estimate a post-RL semantic-occurrence functional and state an explicit residual condition under which this observable quantity captures part of the path-volume effect.

The estimator is organized at three levels. A state-level deviation $o(c)$ measures the relative free-energy displacement of a semantic state c from the most recurrent state. A path-level deviation $O_\eta(\tau)$ accumulates these state deviations along a complete CoT. Finally, an answer-level deviation $\overline{O}_y$ averages the path deviations among trajectories that terminate at answer y. These quantities summarize how the trajectories supporting an answer are distributed around recurrent low-free-energy semantic states. Under a fixed reference partition, the corresponding answer statistic estimates the conditional occurrence functional G_y. The implemented sample-dependent O/C partition introduces an additional clustering-discrepancy term; neither statistic is asserted to recover Q_y without an additional calibration condition.

Appendix B.2 derives the state-level offset from an occurrence-space Gibbs distribution. Appendix B.3 proves fixed-partition concentration around G_y, adds a finite clustering-discrepancy term for the implemented O/C partition, introduces a reward-conditioned residual model for the relation between G_y and Q_y, and bounds the resulting structural, sampling, and clustering errors. Appendix C specifies the construction and aggregation algorithm. The final rule uses only the empirical mean cumulative offset and does not require a reference model or the latent residual parameters.

B.2 Occurrence-Space Relative Free-Energy Offset

We now coarse-grain the sampled trajectories into recurrent semantic occurrences and estimate relative free energy in the induced occurrence space.

Assume that each raw trajectory τ is represented by $L^{\mathrm{raw}}(\tau)$ semantic states in a discrete space C. Let $N_c^{\mathrm{raw}}(\tau)$ be the number of raw occurrences of state c in τ. The population occurrence distribution is

$$\mu(c) = \frac{\mathbb{E}_{\tau\sim\pi}[N_c^{\mathrm{raw}}(\tau)]}{\mathbb{E}_{\tau\sim\pi}[L^{\mathrm{raw}}(\tau)]}. \tag{35}$$

For N trajectories, define $M = \sum_i L_i^{\mathrm{raw}}$ and $n_c = \sum_i N_c^{\mathrm{raw}}(\tau_i)$. Then $\widehat{\mu}(c) = n_c/M$.

On the augmented space $\Omega = \{(\tau, j) : 1 \le j \le L^{\mathrm{raw}}(\tau)\}$, each occurrence inherits its parent trajectory's Gibbs weight. Hence every cluster has a restricted occurrence-space partition function Z_c^{occ}, and

$$\mu(c) = \frac{Z_c^{\mathrm{occ}}}{Z^{\mathrm{occ}}}. \tag{36}$$

Let $c^* = \arg\max_c \mu(c)$, $\widehat{c}^* = \arg\max_c n_c$, and define the occurrence free energy $F_c = -\beta^{-1}\log Z_c^{\mathrm{occ}}$. The population single-step offset is

$$o(c) = -\log\frac{\mu(c)}{\mu(c^*)}. \tag{37}$$

Proposition 6 (Offset is a relative free-energy difference)**.** *For every observed semantic cluster,*

$$o(c) = \log\frac{Z_{c^*}^{\mathrm{occ}}}{Z_c^{\mathrm{occ}}} = \beta(F_c - F_{c^*}). \tag{38}$$

Its empirical plug-in estimate is

$$\widehat{o}(c) = \log\frac{n_{\widehat{c}^*}}{n_c}. \tag{39}$$

Equation 38 identifies relative free energy in the augmented occurrence space.

B.3 Occurrence Estimation and Residualized Score Error

We fix the input, a reference semantic partition C° that is independent of the sampled trajectories, and $\eta \in (0, 1]$. Let $\widetilde{C}^\circ(\tau)$ be the adjacent-deduplicated cluster sequence under this partition, and define the population path offset and its answer-conditioned mean by

$$O^\circ_\eta(\tau) = \sum_{t=1}^{\widetilde{L}(\tau)} \eta^{t-1} o^\circ(\widetilde{c}_t), \qquad G_y = \mathbb{E}_{\pi(\cdot|y)}[O^\circ_\eta(\tau)]. \tag{40}$$

Let q°_y denote the empirical answer-level mean computed on the same trajectories using C°, and let $q_y := \overline{O}_y$ denote the statistic produced by the implemented, sample-dependent O/C partition $\widehat{C}_N$. Their discrepancy is

$$\Delta^{\text{clust}}_{N,y} := |q_y - q^\circ_y|. \tag{41}$$

Under outcome-only reward, conditioning on the final answer cancels the reward tilt for the reference-partition functional, and hence

$$G_y = \mathbb{E}_{\rho(\cdot|y)}[O^\circ_\eta(\tau)], \qquad Q_y = \sum_{\tau \in \mathcal{T}_y} \rho(\tau). \tag{42}$$

Equation (42) does not imply $G_y \propto Q_y$: G_y is a conditional occurrence functional, whereas Q_y is the total reference mass of the answer fiber. We first establish what the reference-partition statistic estimates before accounting for the additional discrepancy introduced by empirical clustering.

B.3.1 Finite-sample occurrence estimation

Let $\tau_1, \ldots, \tau_N$ be independent samples from π, with $K = |C^\circ|$ reference semantic clusters and $J = |\mathcal{Y}|$ sampled answers. For raw occurrence counts, define

$$X_{i,c} = N^{\text{raw}}_c(\tau_i), \quad A_c = \mathbb{E}[X_{i,c}], \quad \widehat{A}_c = \frac{1}{N}\sum_i X_{i,c} = \frac{n_c}{N}. \tag{43}$$

A1: occurrence regularity. Assume $0 \le X_{i,c} \le B_{\text{occ}}$, $\widetilde{L}(\tau) \le L_{\max}$, and $\alpha_{\min} = \min_c A_c > 0$. This condition retains the bounded-offset property while using it only to control estimation of the post-RL functional G_y. Because the normalizing occurrence count cancels, $o^\circ(c) = \log(A_*/A_c)$, where $A_* = \max_c A_c$. Set

$$r_N = B_{\text{occ}}\sqrt{\frac{\log(6K/\delta)}{2N}}, \qquad W_\eta = \sum_{t=1}^{L_{\max}} \eta^{t-1}, \tag{44}$$

$$V_\eta = W_\eta \log\frac{B_{\text{occ}}}{\alpha_{\min}}, \qquad v_N = \sqrt{\frac{\log(6J/\delta)}{2N}}. \tag{45}$$

Then $0 \le O^\circ_\eta(\tau) \le V_\eta$. For $r_N < \alpha_{\min}$ and $n_y > 0$, define

$$\varepsilon_{N,y} = \frac{2W_\eta r_N}{\alpha_{\min} - r_N} + V_\eta\sqrt{\frac{\log(6J/\delta)}{2n_y}}. \tag{46}$$

Finite clustering error. The implemented O/C partition is constructed from the sampled trajectories and need not equal C°. We assume only that its induced answer-level statistic has bounded discrepancy: for some $\kappa_N \ge 0$ and $\delta_{\text{cl}} \in [0, 1)$,

$$\Pr\left(\max_{y: n_y > 0} \Delta^{\text{clust}}_{N,y} \le \kappa_N\right) \ge 1 - \delta_{\text{cl}}. \tag{47}$$

This condition allows imperfect clustering and does not assert that κ_N vanishes universally. It isolates the component of the implemented estimator that is not controlled by fixed-partition concentration.

Theorem 1 (Sampling and clustering error)**.** *Under A1 and Eq.* (47)*, with probability at least* $1 - \delta - \delta_{\mathrm{cl}}$*, simultaneously for every answer with* $n_y > 0$*,*

$$\boxed{|q_y - G_y| \le \widetilde{\varepsilon}_{N,y}, \qquad |\widehat{p}_y - p_y| \le v_N,} \qquad \widehat{p}_y = \frac{n_y}{N}. \tag{48}$$

Here

$$\widetilde{\varepsilon}_{N,y} := \varepsilon_{N,y} + \kappa_N. \tag{49}$$

Without imposing Eq. (47)*, the corresponding pathwise statement is* $|q_y - G_y| \le \varepsilon_{N,y} + \Delta^{\mathrm{clust}}_{N,y}$.

Proof. Hoeffding's inequality and a union bound over clusters imply $\max_c |\widehat{A}_c - A_c| \le r_N$, with failure probability at most $\delta/3$. The maximum is nonexpansive, so $|\max_c \widehat{A}_c - A_*| \le r_N$, even if the empirical modal cluster changes. The mean-value bound for the logarithm therefore gives

$$\max_c |\widehat{o}^{\circ}(c) - o^{\circ}(c)| \le \frac{2r_N}{\alpha_{\min} - r_N}. \tag{50}$$

Accumulation along a compressed path multiplies this bound by at most W_η. Conditional on n_y, the reference-partition population offsets of trajectories ending in y are independent draws from $\pi(\cdot \mid y)$ and lie in $[0, V_\eta]$. A second application of Hoeffding's inequality and a union bound over answers yield

$$\left| \frac{1}{n_y} \sum_{i:g(\tau_i)=y} O^{\circ}_{\eta}(\tau_i) - G_y \right| \le V_\eta \sqrt{\frac{\log(6J/\delta)}{2n_y}} \tag{51}$$

with failure probability at most $\delta/3$. Combining these events gives

$$\begin{aligned} |q^{\circ}_y - G_y| \le \frac{1}{n_y} \sum_{i:g(\tau_i)=y} |\widehat{O}^{\circ}_{\eta}(\tau_i) - O^{\circ}_{\eta}(\tau_i)| \\ + \left| \frac{1}{n_y} \sum_{i:g(\tau_i)=y} O^{\circ}_{\eta}(\tau_i) - G_y \right| \le \varepsilon_{N,y}. \end{aligned} \tag{52}$$

The uniform first event permits occurrence counts and answer means to reuse the same trajectories. Applying Hoeffding's inequality to answer indicators and taking one final union bound gives $|\widehat{p}_y - p_y| \le v_N$ for all answers. Finally,

$$|q_y - G_y| \le |q_y - q^{\circ}_y| + |q^{\circ}_y - G_y| \le \Delta^{\mathrm{clust}}_{N,y} + \varepsilon_{N,y}. \tag{53}$$

Intersecting this event with Eq. (47) completes the proof. □

B.3.2 Reward-conditioned path-volume residual

Theorem 1 separates fixed-partition sampling error from the discrepancy introduced by empirical O/C clustering; neither component by itself identifies reference mass. We use the following weaker calibration condition to characterize how much of the reference path-volume term is captured by G_y.

A2: reward-conditioned residual. For every answer with $G_y, Q_y > 0$, assume

$$\delta_y := \log \frac{G_y}{Q_y} = h_x(r(y)) + \xi_y, \qquad |\xi_y| \le \epsilon_x, \tag{54}$$

where h_x is shared by answers at the same reward level. This assumption does not require a common scale between G_y and Q_y; the answer-specific term ξ_y explicitly records the path-volume component that remains unresolved.

Combining Eq. (54) with the answer Gibbs identity gives the exact decomposition

$$\log \frac{p_y}{G_y} = \Phi_x(r(y)) - \log Z - \xi_y, \qquad \Phi_x(u) := \beta u - h_x(u). \tag{55}$$

Thus normalization by G_y removes the explicit $\log Q_y$ term and converts its remaining answer-specific influence into ξ_y; it does not claim exact recovery of reward or of Q_y.

Proposition 7 (Partial path-volume estimation)**.** *On the event of Theorem 1, suppose* $\widetilde{\varepsilon}_{N,y} < G_y$ *and* $v_N < p_y$*. Then*

$$\boxed{\left|\log q_y - \log Q_y - h_x(r(y))\right| \le \epsilon_x - \log\left(1 - \frac{\widetilde{\varepsilon}_{N,y}}{G_y}\right).} \tag{56}$$

Moreover, for $\widehat{s}_y := \log(\widehat{p}_y/q_y)$*,*

$$\boxed{\left|\widehat{s}_y - \left[\Phi_x(r(y)) - \log Z - \xi_y\right]\right| \le D_{N,y},} \tag{57}$$

where

$$D_{N,y} := -\log\left(1 - \frac{v_N}{p_y}\right) - \log\left(1 - \frac{\widetilde{\varepsilon}_{N,y}}{G_y}\right). \tag{58}$$

Proof. Equation (54) gives $\log G_y = \log Q_y + h_x(r(y)) + \xi_y$. Since $|q_y - G_y| \le \widetilde{\varepsilon}_{N,y} < G_y$,

$$|\log q_y - \log G_y| \le -\log\left(1 - \frac{\widetilde{\varepsilon}_{N,y}}{G_y}\right). \tag{59}$$

The first result follows by the triangle inequality and $|\xi_y| \le \epsilon_x$. Similarly, $|\widehat{p}_y - p_y| \le v_N < p_y$ implies

$$|\log \widehat{p}_y - \log p_y| \le -\log\left(1 - \frac{v_N}{p_y}\right). \tag{60}$$

Combining the two logarithmic deviations with Eq. (55) proves the second result. □

Corollary 1 (Sufficient condition for correct-answer ranking)**.** *Let* $y^\star$ *be the unique rewarded correct answer in the sampled support, and define*

$$m_\Phi := \min_{y \ne y^\star} \left\{\Phi_x(r(y^\star)) - \Phi_x(r(y))\right\}. \tag{61}$$

On the event of Theorem 1, suppose the support conditions of Proposition 7 hold. If

$$m_\Phi > 2\epsilon_x + D_{N,y^\star} + \max_{y \ne y^\star} D_{N,y}, \tag{62}$$

then $y^\star$ *is the unique maximizer of* $\widehat{s}_y$*, and hence the unique minimizer of the implemented score* q_y/n_y*.*

Proof. For every $y \ne y^\star$, Eq. (57) gives

$$\widehat{s}_{y^\star} - \widehat{s}_y \ge \Phi_x(r(y^\star)) - \Phi_x(r(y)) - (\xi_{y^\star} - \xi_y) - D_{N,y^\star} - D_{N,y} \tag{63}$$

$$\ge m_\Phi - 2\epsilon_x - D_{N,y^\star} - D_{N,y} > 0, \tag{64}$$

where the second inequality uses $|\xi_y| \le \epsilon_x$. Therefore no incorrect sampled answer can match or exceed the score of $y^\star$. □

Under fixed positive supports, $D_{N,y} = O_p(N^{-1/2} + \kappa_N)$. If $\kappa_N \to 0$, then $m_\Phi > 2\epsilon_x$ is a population sufficient condition and yields correct selection with probability tending to one as $N \to \infty$. Without vanishing clustering error, the finite-sample condition in Eq. (62) remains valid, but no asymptotic consistency claim follows. The condition makes the unresolved parts explicit: the correction need not recover Q_y, and it preserves the correct ranking only when the reward-conditioned margin exceeds the structural, sampling, and clustering errors.

Since $\widehat{p}_y/q_y = n_y/(Nq_y)$, maximizing $\widehat{s}_y$ is equivalent to minimizing the implemented score q_y/n_y; the common factor N has no effect on the ranking. The algorithm requires neither Q_y, h_x, nor ξ_y; these quantities only characterize the correction represented by its observable score.

Conditionally sub-Gaussian residuals. For a probabilistic refinement of A2, choose h_x to center the residual within each reward level and suppose it satisfies

$$\mathbb{E}\left[e^{\lambda \xi_y} \mid r(y)\right] \le e^{\lambda^2 \sigma_x^2/2} \qquad \text{for all } \lambda \in \mathbb{R}. \tag{65}$$

No independence across answers is required. A two-sided tail bound followed by a union bound gives, with probability at least $1 - \delta_r$,

$$\max_{1 \le y \le J} |\xi_y| \le \sigma_x \sqrt{2 \log \frac{2J}{\delta_r}}. \tag{66}$$

Substituting this radius for ϵ_x in Eq. (56) yields a uniform probabilistic partial-estimation bound. This is a distributional assumption over answer fibers or tasks, not a consequence of the Gibbs identity.

Data-dependent choice of η. If η is selected from a fixed finite grid $\mathcal{H}$, replace $\log(6J/\delta)$ in the answer-conditioned term of Eq. (46) by $\log(6J|\mathcal{H}|/\delta)$. The resulting union bound makes Theorem 1 valid for the selected grid value when Eq. (47) holds uniformly over that grid. The residual parameters may depend on the input and η, but h_x remains shared within a reward level.

B.3.3 Error-rate analysis

When $r_N \le \alpha_{\min}/2$, the explicit sampling radius obeys

$$\varepsilon_{N,y} = O\left(\frac{W_\eta B_{\text{occ}}}{\alpha_{\min}} \sqrt{\frac{\log(K/\delta)}{N}} + V_\eta \sqrt{\frac{\log(J/\delta)}{n_y}}\right). \tag{67}$$

For fixed cluster and answer supports, bounded path and occurrence parameters, and positive lower bounds on $\alpha_{\min}$, p_y, and G_y, we have $n_y = \Theta_p(Np_y)$, and hence

$$\boxed{D_{N,y} = O_p(N^{-1/2} + \kappa_N), \qquad \left|\log q_y - \log Q_y - h_x(r(y))\right| = O(\epsilon_x) + O_p(N^{-1/2} + \kappa_N).} \tag{68}$$

Under Eq. (65), the simultaneous bound over J candidates becomes

$$O\left(\sigma_x \sqrt{\log(J/\delta_r)}\right) + O_p(N^{-1/2} + \kappa_N). \tag{69}$$

Increasing N reduces the fixed-partition sampling component. It does not remove the structural residual, and it removes clustering discrepancy only when $\kappa_N \to 0$. The controlled diagnostic in Appendix D.2 examines when the structural assumption is plausible and where it becomes loose.

C Complete Algorithms

Algorithm 2 gives the complete inference procedure. Algorithms 3–5 then expand its three substantive subroutines.

Algorithm 2 Complete SelFocus inference

Require: Questions $\mathcal{Q}$, sampled CoTs $\mathcal{T}_q$ for each q, encoder E, NLI classifier V, and seed ξ
Ensure: Run-level coefficient η and selected answer sets $\{\widehat{\mathcal{Y}}_q\}_{q \in \mathcal{Q}}$

```
for each q ∈ Q do
    A_q ← BuildSemanticPaths(q, T_q, E, V) using Algorithm 3.
end for
η ← CalibrateDiscount(Q, {A_q}_q, ξ) using Algorithm 4.
for each q ∈ Q do
    Ŷ_q ← AggregateAnswers(q, A_q, η) using Algorithm 5.
end for
return η, {Ŷ_q}_{q∈Q}.
```

C.1 Semantic Occurrence Construction

The preceding estimator assumes a semantic partition of the sampled trajectories. We construct this partition using a fixed two-stage O/C procedure. For each question, every sampled CoT is parsed into an ordered sequence of semantic steps. Stage 1 (O) partitions occurrences by step index and merges semantically overlapping steps only within the same index across trajectories. Stage 2 (C) compares the resulting Stage 1 clusters across step indices and merges them when their content overlaps or one semantically contains the other. Embedding retrieval proposes high-recall candidate pairs, and bidirectional NLI determines whether a merge is admissible. The fixed NLI classifier V judges semantic relations between steps, not the correctness of a derivation or final answer, and is not trained as part of Selfocus. Retrieval therefore affects computational efficiency but not the merge criterion. The same O/C relation sequence is used for every model and benchmark.

For question q, let $c_{q,i,t}$ denote the final cluster assigned to raw step t of trajectory $\tau_{q,i}$. We count occurrences within the question and before any path compression,

$$n_{q,c} = \sum_{i=1}^{N_q} \sum_{t=1}^{L_{q,i}^{\text{raw}}} \mathbb{I}\{c_{q,i,t} = c\}, \qquad \widehat{c}_q^{\,*} \in \arg\max_c n_{q,c}, \qquad \widehat{o}_q(c) = \log \frac{n_{q,\widehat{c}_q^{\,*}}}{n_{q,c}}. \tag{70}$$

Thus repeated visits remain evidence for estimating semantic occurrence mass, but no occurrence count is shared across questions. Adjacent repeated cluster IDs are removed only when converting a trajectory into a path-level offset.

Algorithm 3 Question-level semantic occurrence construction

Require: Question q, sampled CoTs $\mathcal{T}_q$, encoder E, and bidirectional NLI classifier V
Ensure: Answers $y_{q,i}$, compressed paths $\widetilde{C}_{q,i}$, and offsets $\widehat{o}_q(c)$

for each $\tau_{q,i} \in \mathcal{T}_q$ **do**
  Parse final answer $y_{q,i}$ and ordered raw steps $(s_{q,i,1}, \ldots, s_{q,i,L_{q,i}^{\text{raw}}})$.
end for
Initialize every raw step as a singleton cluster.
for each step index t **do**
  Retrieve candidate pairs only within $\{s_{q,i,t}\}_i$ using E.
  for each same-index candidate pair (s_a, s_b) **do**
    if V identifies bidirectional semantic overlap **then**
      Merge the clusters containing s_a and s_b. *Stage 1: O*
    end if
  end for
end for
Retrieve candidate pairs across step indices among the Stage 1 clusters.
for each cross-index candidate cluster pair (C_a, C_b) **do**
  if V identifies bidirectional overlap or semantic containment **then**
    Merge C_a and C_b. *Stage 2: C*
  end if
end for
Assign final cluster IDs $c_{q,i,t}$ and compute question-local $\widehat{o}_q(c)$ by Eq. (70).
for each $\tau_{q,i} \in \mathcal{T}_q$ **do**
  $\widetilde{C}_{q,i} \leftarrow \text{AdjacentDedup}(c_{q,i,1}, \ldots, c_{q,i,L_{q,i}^{\text{raw}}})$.
end for
return $\{y_{q,i}, \widetilde{C}_{q,i}\}_i, \{\widehat{o}_q(c)\}_c$.

C.2 Unsupervised Estimation of the Step-Discount Coefficient

The single-step offsets determine how far each semantic state lies from the modal state, but they do not determine how evidence should be accumulated across reasoning depth. We estimate one discount coefficient per experimental run from an unlabeled calibration subset. Let Q be the questions in the run and sample uniformly without replacement with a fixed seed,

$$C \subset Q, \qquad |C| = \lceil 0.3|Q| \rceil . \tag{71}$$

Every calibration question uses all of its sampled trajectories. The procedure uses neither extracted answers nor correctness labels, and statistics are never pooled across questions or runs.

For a compressed path $\widetilde{C}_{q,i} = (c_{q,i,1}, \ldots, c_{q,i,\widetilde{L}_{q,i}})$, define the occupancy of cluster c at position t, the number of paths reaching that position, and the number of singleton nodes as

$$m_{q,t,c} = \sum_i \mathbb{I}\{t \le \widetilde{L}_{q,i},\ c_{q,i,t} = c\}, \qquad N_{q,t} = \sum_c m_{q,t,c}, \qquad F_{q,t} = \sum_c \mathbb{I}\{m_{q,t,c} = 1\}. \tag{72}$$

The post-initial repeated-support rate is

$$R_q = 1 - \frac{\sum_{t\ge 2} F_{q,t}}{\sum_{t\ge 2} N_{q,t}}. \tag{73}$$

Thus a later node is supported only when another path reaches the same O/C cluster at the same compressed position.

We separately measure whether paths become less central with depth. For each path with $\widetilde{L}_{q,i} \ge 2$, let $h_{q,i} = \lfloor \widetilde{L}_{q,i}/2 \rfloor$ and define

$$d_{q,i} = \frac{1}{h_{q,i}} \left(\sum_{t=\widetilde{L}_{q,i}-h_{q,i}+1}^{\widetilde{L}_{q,i}} \widehat{o}_q(c_{q,i,t}) - \sum_{t=1}^{h_{q,i}} \widehat{o}_q(c_{q,i,t}) \right). \tag{74}$$

The middle state of an odd-length path is excluded. Let $\mathcal{P}_q^+$ contain the paths of length at least two and set

$$D_q = \begin{cases} |\mathcal{P}_q^+|^{-1} \sum_{i\in\mathcal{P}_q^+} d_{q,i}, & |\mathcal{P}_q^+| > 0, \\ 0, & |\mathcal{P}_q^+| = 0. \end{cases} \tag{75}$$

Writing C_R for calibration questions with at least one post-initial node, we aggregate questions with equal weight whenever $C_R \ne \varnothing$:

$$R = \frac{1}{|C_R|} \sum_{q\in C_R} R_q, \qquad D = \frac{1}{|C|} \sum_{q\in C} D_q. \tag{76}$$

Finally,

$$\eta_{\text{raw}} = R\exp(-D), \qquad \eta = \Pi_{\mathcal{E}}\Big(\text{clip}_{[0.1,1]}(\eta_{\text{raw}})\Big), \qquad \mathcal{E} = \{0.1, 0.2, \ldots, 1.0\}, \tag{77}$$

where $\Pi_{\mathcal{E}}$ selects the nearest grid value and maps an exact midpoint to the smaller value. If no calibration path has length at least two, we set $\eta = 1$. A lower repeated-support rate or a larger increase in late-path offset therefore produces stronger discounting. This coefficient is a label-free structural heuristic, not an unbiased estimate or a consequence of the free-energy identity.

Algorithm 4 Unlabeled run-level calibration of η

Require: Questions Q, semantic paths and offsets from Algorithm 3, and seed ξ
Ensure: Run-level discount coefficient η

$C \leftarrow \text{UniformSample}(Q, \lceil 0.3|Q| \rceil; \xi)$.
if no path in C has compressed length at least two **then**
  return $\eta = 1$.
end if
Initialize $C_R \leftarrow \varnothing$.
for each $q \in C$ **do**
  Compute $m_{q,t,c}$, $N_{q,t}$, and $F_{q,t}$ by Eq. (72).
  if $\sum_{t\ge 2} N_{q,t} > 0$ **then**
    Compute R_q by Eq. (73); add q to C_R.
  end if
  Form $\mathcal{P}_q^+ = \{i : \widetilde{L}_{q,i} \ge 2\}$.
  Compute $d_{q,i}$ for $i \in \mathcal{P}_q^+$ and D_q by Eqs. (74)–(75).
end for
Compute the question-balanced R and D by Eq. (76).
$\eta_{\text{raw}} \leftarrow R\exp(-D)$.
$\eta \leftarrow \Pi_{\mathcal{E}}(\text{clip}_{[0.1,1]}(\eta_{\text{raw}}))$, mapping an exact grid midpoint to the smaller value.
return η.

C.3 Path and Answer Aggregation

After estimating η, we hold it fixed for the complete run. For a compressed path $\widetilde{C}(\tau) = (c_1, \ldots, c_{\widetilde{L}})$, its cumulative offset is

$$\widehat{O}_{\eta,q}(\tau) = \sum_{t=1}^{\widetilde{L}} \eta^{t-1} \widehat{o}_q(c_t), \qquad 0 < \eta \le 1. \tag{78}$$

At $\eta = 1$, this is a negative log composite-likelihood ratio to the modal semantic region. For $\eta < 1$, the estimated coefficient geometrically discounts later states whose recurrence is weaker or whose offsets drift upward. The exact free-energy interpretation applies to the undiscounted offsets; depth weighting is the subsequent aggregation rule.

For answer y, let $\mathcal{I}_y$ contain its scorable paths and let $n_y = |\mathcal{I}_y|$ be its vote count. We first average path offsets,

$$\overline{O}_y = \frac{1}{|\mathcal{I}_y|} \sum_{i \in \mathcal{I}_y} \widehat{O}_{\eta,q}(\tau_i). \tag{79}$$

and then divide by its vote count:

$$S(y) = \frac{\overline{O}_y}{n_y}, \qquad \widehat{\mathcal{Y}} = \arg\min_{y \in \mathcal{Y}_q} S(y), \tag{80}$$

where $\mathcal{Y}_q$ is the set of distinct normalized answers sampled for question q. This is the answer rule used throughout the paper.

Algorithm 5 Question-level path and answer aggregation

Require: Question q, outputs $\{y_{q,i}, \widetilde{C}_{q,i}\}_i$, $\{\widehat{o}_q(c)\}_c$ from Algorithm 3, and η
Ensure: Selected answer set $\widehat{\mathcal{Y}}_q$

```
for each scorable trajectory τ_{q,i} do
    Compute Ô_{η,q}(τ_{q,i}) by Eq. (78).
end for
for each distinct answer y do
    I_{q,y} ← {i : y_{q,i} = y} and n_{q,y} ← |I_{q,y}|.
    Ō_{q,y} ← n_{q,y}^{-1} Σ_{i∈I_{q,y}} Ô_{η,q}(τ_{q,i}).
    S_q(y) ← Ō_{q,y} / n_{q,y}.
end for
Ŷ_q ← arg min_{y∈Y_q} S_q(y).
return Ŷ_q.
```

The overview in Algorithm 2 makes the scope of each statistic explicit. Algorithm 3 estimates question-local semantic recurrence and occurrence-space offsets. Algorithm 4 uses only the sampled path structure of the unlabeled calibration questions to obtain one run-level η. Algorithm 5 then applies that fixed coefficient to every question and combines trajectory deviation with endpoint frequency. Together, the O/C relation sequence, count-before-compression convention, coefficient estimator, and score in Eq. (80) constitute the complete inference rule.

D Experimental Setup and Reproducibility

D.1 Controlled Experiment Setup

The controlled experiment in Section 3.2 uses a six-step continuous control problem. The state begins at $s_0 = 0 \in \mathbb{R}^d$, the terminal target is $s_T^* = e_1$, and each action is an additive increment,

$$a_t \sim \mathcal{N}\Big(\mu_\theta(s_t, t), \sigma_t^2 I_d\Big), \qquad s_{t+1} = s_t + a_t. \tag{81}$$

The policy mean is a two-hidden-layer MLP with tanh activations and hidden width $\max(48, 2d)$; its diagonal log standard deviation is learned during training. A separate policy is trained for every $d \in \{2, 4, 8, 16\}$ by REINFORCE for 2,500 updates, using batches of 1,024 trajectories, Adam with learning rate 3×10^{-3}, initial standard deviation 0.18, and terminal reward

$$R(\tau) = -\frac{\|s_T - e_1\|_2^2}{2(0.35)^2}. \tag{82}$$

The training baseline is an exponential moving average with coefficient 0.98. All experiments use seed 7.

For the dimension sweep, each trained policy is evaluated with fixed per-coordinate sampling standard deviation $\sigma_{\text{eval}} = 0.035$ and 60,000 trajectories. For the dispersion sweep, the learned $d = 2$ mean policy is frozen and evaluated at $\sigma_{\text{eval}} \in \{0.035, 0.070, 0.140, 0.280\}$. Trajectory differential entropy is estimated from the regularized empirical covariance of the concatenated six-step state sequence. These two interventions separately vary effective trajectory dimension and local stochasticity while leaving the reward and terminal target unchanged.

D.2 Reward-Conditioned Residual Diagnostic

We use the same frozen multi-step policies to test A2 in a setting where reference answer mass is observable. Let $\bar{s}_t$ denote the deterministic state at step t, obtained by rolling out the learned mean policy without sampling noise. For a sampled trajectory, the continuous step deviation and its normalized radius are

$$d_t(\tau) = s_t(\tau) - \bar{s}_t, \qquad \nu_t(\tau) = \frac{\|d_t(\tau)\|_2}{\sigma_{\text{eval}}\sqrt{td}}. \tag{83}$$

We discretize ν_t using the fixed radius edges $0.45, 0.85, 1.30, 2.0, 3.0$. The innermost interval is one central semantic state. In every other interval, the direction $d_t/\|d_t\|_2$ is assigned to the most similar of eight fixed unit direction prototypes by maximum inner product. The radius edges and prototypes are fixed before computing any answer-level quantity, so this diagnostic has no clustering discrepancy.

The terminal answer map is also fixed before estimation: one radius-0.1 correct ball and 64 mutually exclusive wrong regions formed by four radial shells and 16 directional sectors. We repeat the directional partition with eight fixed seeds. A reference split of one million trajectories is sampled from the frozen Gaussian policy ρ, and

$$\widehat{Q}_y = \frac{1}{N_\rho} \sum_{i=1}^{N_\rho} \mathbf{1}\{g(\tau_i) = y\} \tag{84}$$

estimates the un-tilted reference mass. To estimate post-RL semantic occurrence mass, the same reference trajectories receive importance weights

$$w_i = \frac{\exp\{\beta r(g(\tau_i))\}}{\sum_{j=1}^{N_\rho} \exp\{\beta r(g(\tau_j))\}}, \qquad \beta = 2, \tag{85}$$

and cluster occurrence counts are weighted by w_i. These weights are the Monte Carlo Gibbs tilt from ρ to π, so the resulting occurrence offsets correspond to the post-RL policy.

A separate one-million-trajectory split, independently sampled from ρ, is mapped to the fixed semantic states and answers. Using the offsets from Eq. (85), we estimate

$$\widehat{G}_y = \frac{1}{N_y} \sum_{i:g(\tau_i)=y} O_\eta(\tau_i), \qquad \eta \in \{0.5, 1\}. \tag{86}$$

This independent split may be sampled from ρ because the reward is outcome-only. For all trajectories ending at the same answer, the Gibbs factor is constant and cancels after conditioning; hence

$$\pi(\tau \mid y) = \rho(\tau \mid y). \tag{87}$$

Thus $\widehat{Q}_y$ estimates reference answer mass, the importance-weighted occurrence counts estimate the semantic offsets under π, and $\widehat{G}_y$ is evaluated independently under the correct answer-conditional trajectory distribution.

For the bounded-residual version of A2, $h_x(r)$ is fitted as the minimax center of $\delta_y = \log(G_y/Q_y)$ within each reward level, and ϵ_x is the largest absolute residual around that center. We evaluate the sub-Gaussian refinement separately: even directional sectors form a calibration split and odd sectors form a held-out split, and δ_y is centered by its reward-level mean within each split. We then compute, over a fixed grid of λ values, the smallest empirical σ satisfying $\log \mathbb{E}[\exp(\lambda\delta_y)] \le \lambda^2\sigma^2/2$. The inflation ratio is $\sigma_{\text{held}}/\sigma_{\text{cal}}$.

Table 2: **Controlled diagnostic of the reward-conditioned residual model.** Values are means over eight fixed directional partitions. "Supported" requires all 65 answer cells to contain at least 200 trajectories in each independent split.

d	η	Supported	ϵ_x	σ_{cal}	σ_{held}	Inflation
4	0.5	8/8	0.862	0.373	0.359	0.958
4	1.0	8/8	0.846	0.362	0.348	0.958
8	0.5	8/8	0.850	0.409	0.413	1.011
8	1.0	8/8	0.826	0.363	0.367	1.010
16	0.5	8/8	1.955	1.364	1.372	1.006
16	1.0	8/8	1.926	1.337	1.345	1.006

The held-out MGF scale closely tracks the calibration scale in all supported conditions, with mean inflation between 0.958 and 1.011. This supports the plausibility of a reward-conditioned sub-Gaussian tail shape inside the controlled simulator. The absolute scale is not dimension independent: σ is approximately 0.35–0.41 for $d = 4, 8$ but increases to approximately 1.34–1.37 for $d = 16$, while the uniform residual radius increases to about 1.94. The $d = 2$ condition is excluded from the primary table because the correct ball contains nearly all reference mass and not all wrong cells meet the support threshold. This split diagnostic evaluates residual scale after separately centering the calibration and held-out partitions. It therefore neither tests transfer of a single fitted h_x nor directly establishes an improvement in answer ordering.

Corollary 1 converts the residual model into a directly falsifiable ranking statement. Table 3 evaluates its population form, $m_\Phi > 2\epsilon_x$, where the answer partition is fixed in advance, the clustering discrepancy is zero, and the finite-sample radii vanish because Q_y, G_y, and p_y are estimated from the independent million-trajectory populations. The exact corrected margin is

$$M_{\text{exact}} := \min_{y \neq y^\star} \left\{ \log \frac{p_{y^\star}}{G_{y^\star}} - \log \frac{p_y}{G_y} \right\}. \tag{88}$$

Table 3: **Controlled test of the correct-ranking condition.** Values are means over eight fixed partitions. "Generic pass" is the percentage satisfying $m_\Phi > 2\epsilon_x$; "exact pass" is the percentage with $M_{\text{exact}} > 0$.

d	η	m_Φ	ϵ_x	$m_\Phi - 2\epsilon_x$	Generic pass	M_{exact}	Exact pass
4	0.5	8.616	0.862	6.892	100.0%	7.754	100.0%
4	1.0	8.760	0.846	7.068	100.0%	7.914	100.0%
8	0.5	6.416	0.850	4.717	100.0%	5.567	100.0%
8	1.0	6.570	0.826	4.917	100.0%	5.744	100.0%
16	0.5	3.675	1.955	−0.234	0.0%	1.720	100.0%
16	1.0	3.703	1.926	−0.149	12.5%	1.777	100.0%

The generic sufficient condition holds in every partition for four of the six supported (d, η) settings. It is conservative: the exact corrected margin remains positive in every partition for all six settings. At $d = 16$, the uniform radius is governed by the most extreme answer-specific residual, so the generic condition can fail even though the realized pairwise margins remain positive. This experiment therefore verifies a concrete algorithmic implication of the residual model in the controlled simulator while also exposing where its uniform sufficient bound becomes loose.

Direct ranking test. We separately test the ordering consequence without fitting or recentering h_x. We retain the frozen policies, $\beta = 2$, and the four-shell by 16-sector partition of the wrong-answer

region. For each $d \in \{4, 8, 16\}$, we vary the reference mass of the centered correct region from 0.75% to 1.50% in increments of 0.05 percentage points. The corresponding radius is fixed as a quantile of one million reference trajectories before evaluating any selector. These trajectories estimate Q_y, and the Gibbs importance weights in Eq. (85) estimate the post-RL occurrence offsets. An independent set of one million trajectories estimates the answer-conditioned mean cumulative offset G_y using Eq. (87). We repeat eight fixed directional partitions.

For the unique rewarded answer $y^\star$, the post-RL answer distribution is constructed exactly as

$$p_y = \frac{Q_y \exp(\beta \mathbf{1}\{y = y^\star\})}{\sum_z Q_z \exp(\beta \mathbf{1}\{z = y^\star\})}. \tag{89}$$

Vote ranks answers by p_y, population SELFOCUS with $\eta = 1$ ranks by p_y/G_y, equivalently minimizing G_y/p_y, and the strict oracle ranks by p_y/Q_y. We report the fraction of partitions in which $y^\star$ is the unique Top-1 answer and its mean rank across partitions.

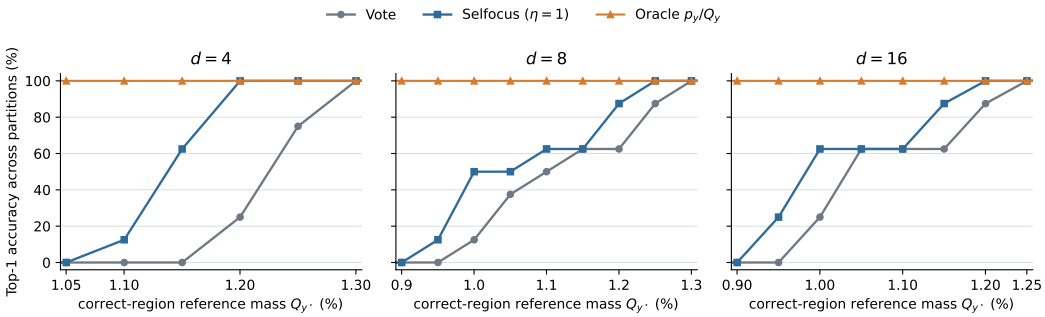


Figure 6: **Direct comparison of Vote, population SELFOCUS, and the strict oracle.** Top-1 accuracy is measured over eight fixed answer partitions. Each panel displays the informative transition from the final zero-accuracy SELFOCUS point to the first 100%-accuracy Vote point; the complete sweep uses $Q_{y^\star} = 0.75\%$–1.50% at 0.05-percentage-point resolution.

Table 4: **Representative regimes from the direct ranking test.** Each method entry is Top-1 accuracy (%) / mean rank of the uniquely rewarded answer over eight fixed partitions. Lower rank is better.

d	$Q_{y^\star}$ (%)	Vote	SELFOCUS	Oracle p_y/Q_y	Regime
4	1.05	0.0 / 3.875	0.0 / 2.750	100.0 / 1.000	shared failure
4	1.15	0.0 / 2.500	62.5 / 1.375	100.0 / 1.000	partial recovery
4	1.20	25.0 / 1.750	100.0 / 1.000	100.0 / 1.000	full recovery
8	0.90	0.0 / 5.375	0.0 / 4.250	100.0 / 1.000	shared failure
8	1.00	12.5 / 2.875	50.0 / 2.000	100.0 / 1.000	partial recovery
8	1.20	62.5 / 1.500	87.5 / 1.125	100.0 / 1.000	partial recovery
16	0.90	0.0 / 5.125	0.0 / 4.000	100.0 / 1.000	shared failure
16	1.00	25.0 / 3.000	62.5 / 1.625	100.0 / 1.000	partial recovery
16	1.15	62.5 / 1.375	87.5 / 1.125	100.0 / 1.000	partial recovery

The oracle is Top-1 in every condition because Eq. 89 gives $p_y/Q_y \propto \exp(\beta r(y))$; it is therefore a construction check rather than a learned selector. In the transition region, the SELFOCUS curve moves consistently ahead of Vote. The largest separation occurs at $d = 4$ and $Q_{y^\star} = 1.20\%$, where Vote is correct in 25.0% of partitions and SELFOCUS in 100.0%. At $d = 8$ and $Q_{y^\star} = 1.00\%$, the corresponding accuracies are 12.5% and 50.0%, and the correct answer's mean rank improves from 2.875 to 2.000. At $d = 16$ and the same mass, accuracy rises from 25.0% to 62.5%, while mean rank improves from 3.000 to 1.625.

The correction is partial. At $d = 4, Q_{y^\star} = 1.05\%$ and at $d \in \{8, 16\}, Q_{y^\star} = 0.90\%$, both Vote and SELFOCUS have zero Top-1 accuracy, although SELFOCUS still improves the mean rank of the rewarded answer. Thus, within this controlled construction where the Gibbs model and A2 can be evaluated directly, the observable occurrence statistic reduces path-volume-induced ranking bias over

a nontrivial transition interval. It neither identifies Q_y nor guarantees recovery when the rewarded answer region is too sparse, and this simulator result does not establish the mechanism behind the benchmark gains.

D.3 Models and Benchmarks

We evaluate four open-weight instruction or reasoning models: Gemma-2-9B-IT, Llama-3.3-70B-Instruct, Qwen3-32B, and GPT-OSS-120B. Table 5 gives the benchmark inventory. TheoremQA is restricted to its 747 text-only questions so that all evaluated inputs are textual.

Table 5: **Benchmark inventory.** "Open" denotes a free-form answer that is canonicalized before aggregation.

Benchmark	Questions	Answer format	Evaluated models
GPQA Diamond	198	Four-way choice	Gemma, Llama, Qwen
MATH500	500	Open mathematical	Gemma, Llama, Qwen
TheoremQA	747	Typed open answer	Gemma, Llama, Qwen
AIME25	30	Integer	Llama, Qwen, GPT-OSS
AIME26	30	Integer	Llama, Qwen, GPT-OSS
BRUMO25	30	Integer	GPT-OSS
HMMT25	30	Integer	GPT-OSS

For every model–dataset cell, the target sampling budget is 64 independently decoded CoTs per question. Each experiment is reported over four independent runs.

D.4 Unified Prompt and Generation Protocol

All main experiments use the same clarify-boundary instruction. Multiple-choice questions use the following user-message template:

```
{question}
Options:
A. {option_A}
B. {option_B}
C. {option_C}
D. {option_D}
Please reason step by step and clarify the boundary between steps,
and put your final answer within \boxed{}.
```

Open-answer questions use:

```
{question}
Please reason step by step and clarify the boundary between steps,
and put your final answer within \boxed{}.
```

The user message is wrapped with each checkpoint's native chat template. Independent stochastic decoding is used to obtain 64 outputs. Table 6 gives the complete decoding configuration. Top-k truncation is disabled for these generations, and SelFocus does not request token log probabilities.

Table 6: **Generation configuration by model and benchmark.** "Standard" denotes the checkpoint's ordinary instruction mode; Qwen thinking and GPT-OSS reasoning effort are stated explicitly.

Model	Benchmarks	Mode	Temp.	Top-p	Max output
Gemma2-9B	GPQA	Standard	0.7	0.90	4,096
	MATH500, TheoremQA	Standard	0.7	0.96	4,096
Llama3.3-70B	GPQA	Standard	0.7	0.90	4,096
	MATH500, TheoremQA	Standard	0.7	0.96	4,096
	AIME25, AIME26	Standard	0.7	0.96	32,768
Qwen3-32B	GPQA	Non-thinking	0.7	0.90	8,192
	MATH500, TheoremQA	Non-thinking	0.7	0.96	4,096
	AIME25, AIME26	Thinking	0.7	0.96	32,768
GPT-OSS-120B	AIME25/26, BRUMO25, HMMT25	Low reasoning	0.7	0.96	16,384

D.5 Semantic Path Construction

The clarify-boundary instruction asks the model to expose the internal structure of each CoT. We deterministically split the generated reasoning body at explicit line-initial markers: `Step k` headings, option or choice headings, numbered items, Markdown headings, and standalone bold headings. Markers inside fenced code or display-math blocks are ignored, and divider-only lines do not create steps. Text from one recognized marker up to the next forms one raw step. A preamble before the first marker is retained when it contains at least 12 tokens; if no marker is detected, the remaining reasoning body is treated as one step. The final-answer tail associated with the last `\boxed{}` is excluded, and structural heading markers are removed before semantic comparison.

Every CoT is thereby represented as an ordered sequence of semantic steps and a canonical final answer. The semantic encoder is *sentence-transformers/all-MiniLM-L6-v2*; candidate retrieval uses cosine threshold 0.80 and at most 20 nearest neighbors per step. Semantic relations are then classified by *cross-encoder/nli-deberta-v3-base*, with entailment threshold 0.80 and maximum input length 512 tokens for the concatenated step pair, including special tokens. We checked all tokenized NLI pairs and found that none exceeded this limit; consequently, no NLI input was truncated. This NLI model is a relation classifier: it determines overlap and containment between two reasoning steps and never receives a reference answer or predicts whether a trajectory is correct.

The extracted steps are indexed by their order within each CoT. Stage 1 (O) clusters semantically overlapping occurrences only among steps with the same index across trajectories. All candidate pairs are adjudicated before clusters are formed as connected components of the accepted relation graph. The representative text of each component is its embedding medoid, namely the member with the largest mean cosine similarity to the other members; when similarities are equal, the earliest member in deterministic order is chosen. Stage 1 is completed before Stage 2 begins. Stage 2 (C) compares these Stage 1 medoids across different step indices and merges components related by semantic overlap or one-way semantic containment. It again forms connected components after all accepted edges have been added, so the result is independent of pair-processing order; component identifiers follow the smallest constituent identifier. The O/C construction is fixed across all models and benchmarks. For a raw clustered path $(c_1, \ldots, c_L)$, occurrence counts n_c are computed before compression. Only adjacent repeated identifiers are then removed to form $\widetilde{C}(\tau)$; a later return to the same state remains a distinct path occurrence. The clustered object used in the answer-level derivation is therefore

$$m = \big(\widetilde{C}(\tau), g(\tau)\big), \tag{90}$$

which includes both the compressed semantic trajectory and its final answer.

D.6 Answer Processing and Evaluation

Answers are canonicalized before vote buckets are formed. GPQA outputs are mapped to option letters; AIME-family outputs are mapped to integers; MATH500 and TheoremQA use deterministic type-aware normalization for numerical values, symbolic expressions, sets, intervals, and short textual answers.

All reported SELFOCUS predictions use $S(y) = \overline{O}_y/n_y$ and minimize this score over all distinct normalized answers. The main accuracy table uses the η values returned by the unsupervised, label-free estimator in Eq. (77). Appendix F separately evaluates fixed $\eta = 1$ and the component variants. Accuracy is averaged over four independent runs, and every reported ± value is their sample standard deviation in percentage points.

Baseline implementations and candidate pools. DeepConf is evaluated in its offline form. Each trajectory is assigned its Lowest Group Confidence, defined as the minimum mean token confidence over overlapping 2,048-token windows. We retain the highest-confidence 10% of trajectories and apply confidence-weighted answer aggregation.

For Self-Certainty, trajectory i receives

$$c_i = \frac{1}{T_i} \sum_{t=1}^{T_i} \mathrm{KL}(U \| p_{i,t}). \tag{91}$$

Trajectories are sorted by decreasing c_i. A trajectory at one-indexed rank r receives Borda weight $(N - r + 1)^{1.2}$, and the weights of trajectories with the same normalized answer are summed. For PiCSAR, each decoded reasoning trace r_i is scored by $\log p(r_i \mid x)$. We then append the method's explicit answer instruction $\langle a \rangle$ and teacher-force the answer y_i extracted from that trace, yielding the joint score $\log p(r_i \mid x) + \log p(y_i \mid \langle a \rangle, r_i, x)$. PiCSAR returns the answer of the highest-scoring trajectory.

For Semantic Self-Consistency, we use the semantic-consensus weighting variant (Knappe et al., 2025). Each complete CoT, including its final-answer text, is encoded independently with `Alibaba-NLP/gte-Qwen2-1.5B-instruct` (Li et al., 2023). We use the model's symmetric sentence-similarity interface without a query instruction, its native 32,768-token context limit and last-token pooling, and L_2-normalized 1,536-dimensional embeddings. For trajectory embedding e_i, the trajectory and answer scores are

$$w_i = \sum_{j=1}^{N} \cos(e_i, e_j), \qquad S_{\mathrm{SSC}}(y) = \sum_{i:g(\tau_i)=y} w_i,$$

where the first sum includes $j = i$. We select the answer with the largest $S_{\mathrm{SSC}}(y)$, without clipping or thresholding the cosine similarities. Within every locally recomputed model–dataset cell, Vote and all available selectors reuse the same fixed decoded candidate pool; selector evaluation does not generate or replace trajectories.

E ADDITIONAL EMPIRICAL ANALYSIS

E.1 TRAJECTORY AND ANSWER-CANDIDATE SIGNAL

For a compressed path, mean step offset is its cumulative undiscounted offset divided by compressed length. Normalized answer semantic concentration (ASC) is the mean ratio n_c/n_{c^*} over the uncompressed visited states. These two statistics express the same recurrence signal in opposite directions: recurrent paths have lower offset and higher ASC. Table 7(a) gives the exact model-level means underlying Figure 3; the correct class is shifted toward lower offset and higher normalized ASC for every model family.

At the answer-candidate level, we compare each correct candidate with every incorrect candidate from the same question-run. Offset pairwise accuracy is the probability that the correct candidate has the lower offset; ASC pairwise accuracy is the probability that it has the higher concentration. Table 7(b) shows that the signal generally survives answer projection. In particular, both statistics exceed chance for Gemma, Llama, and Qwen, while normalized ASC remains informative for GPT-OSS. The result supports the use of cross-trajectory recurrence as an answer-selection feature rather than only as a descriptive property of individual CoTs.

Table 7: **Correct-versus-incorrect trajectory and answer-candidate statistics.** Panel (a) reports model-level CoT statistics; lower offset and higher ASC indicate greater semantic recurrence. Panel (b) reports within-question answer-candidate pairwise accuracy, where 0.5 is chance.

	Mean step offset		Normalized ASC	
Model	Correct	Incorrect	Correct	Incorrect
Gemma2-9B	1.342	1.760	0.722	0.684
Llama3.3-70B	1.894	2.395	0.778	0.769
Qwen3-32B	1.370	2.029	0.754	0.732
GPT-OSS-120B	2.191	2.683	0.765	0.723

(b) Answer-candidate discrimination

Model	Offset pairwise	ASC pairwise
Gemma2-9B	0.587	0.602
Llama3.3-70B	0.646	0.626
Qwen3-32B	0.675	0.678
GPT-OSS-120B	0.491	0.639

E.2 Paired Prediction Transitions

Table 8 expands the transition analysis summarized by Figure 4. Across the full evaluation, the dominant off-diagonal transition is from an incorrect Vote prediction to a correct SELFOCUS prediction. On the largest open-answer benchmarks, correction rates substantially exceed regression rates, while most predictions remain unchanged. This explains why the main-table gains are not produced by broadly replacing majority decisions; they arise from a comparatively small proportion of targeted reversals supported by different trajectory structure.

Table 8: **Paired question-level transitions from Vote to SELFOCUS.** All entries are percentages of evaluated question-runs. "Corrected" denotes Vote-incorrect/SELFOCUS-correct transitions, "regressed" denotes the reverse, and Δ is the net accuracy change in percentage points.

Data	Vote	SELFOCUS	Corr.	Regr.	Δ
Gemma2-9B					
GPQA	37.12	39.39	2.78	0.51	+2.27
MATH500	54.70	55.50	2.00	1.20	+0.80
TheoremQA	40.36	42.97	3.68	1.07	+2.61
Llama3.3-70B					
GPQA	47.73	48.74	1.52	0.51	+1.01
MATH500	75.50	76.40	1.00	0.10	+0.90
TheoremQA	65.06	69.21	4.62	0.47	+4.15
AIME25	10.00	13.33	3.33	0.00	+3.33
AIME26	16.67	18.33	1.67	0.00	+1.67

Data	Vote	SELFOCUS	Corr.	Regr.	Δ
Qwen3-32B					
GPQA	59.09	62.37	4.55	1.26	+3.28
MATH500	86.20	86.50	0.50	0.20	+0.30
TheoremQA	77.51	81.73	4.82	0.60	+4.22
AIME25	80.00	80.00	0.00	0.00	0.00
AIME26	85.00	87.50	2.50	0.00	+2.50
GPT-OSS-120B					
AIME25	81.67	85.00	3.33	0.00	+3.33
AIME26	93.33	95.00	1.67	0.00	+1.67
BRUMO25	78.33	75.00	1.67	5.00	-3.33
HMMT25	61.67	66.67	10.00	5.00	+5.00

F Full Ablation Results

Controlled protocol. We isolate the contributions of the score and clustering pipeline on the same independently sampled traces. The ablation contains 17 model–dataset cells, and accuracy is averaged over repeated runs. All variants share the same trajectories, retrieval results, and NLI judgments. All valid answers compete without top-k or top-p filtering. We evaluate both $\eta = 0.5$ and $\eta = 1$ in every cell.

Score decomposition. Starting from the full score $\overline{O}_y / n_y$, we test two modifications. *Pure cumulative offset* removes the answer-frequency divisor n_y. *Mean offset/votes* normalizes each path's cumulative offset by its total position weight before answer averaging while retaining the answer-frequency divisor. We also construct two direct semantic-concentration baselines from each trajectory's mean normalized raw-step cluster occupancy. *Pure ASC* averages this statistic within an answer, whereas

ASC-weighted Vote sums it over all supporting trajectories:

$$S_{\text{Pure-ASC}}(y) = \frac{1}{n_y} \sum_{\tau: g(\tau)=y} \text{ASC}(\tau), \qquad S_{\text{ASC-WV}}(y) = \sum_{\tau: g(\tau)=y} \text{ASC}(\tau). \tag{92}$$

Both selectors maximize their respective scores. Their relation, $S_{\text{ASC-WV}}(y) = n_y S_{\text{Pure-ASC}}(y)$, isolates the contribution of answer support while using exactly the same trajectory-level semantic statistic. Neither baseline uses the log-offset transformation or positional discount.

Table 9: **Score-component ablation by model and dataset.** Accuracy (%) is averaged over repeated runs. Pure ASC averages trajectory ASC within each answer; ASC-WV sums it over the answer's supporting trajectories. Both are independent of η.

	Baselines			$\eta = 0.5$			$\eta = 1$		
Dataset	Vote	Pure ASC	ASC-WV	Pure	Mean	Full	Pure	Mean	Full
Gemma2-9B									
GPQA	37.12	27.27	35.86	28.27	39.13	39.13	31.30	37.87	36.86
MATH500	54.70	33.20	53.80	36.30	55.50	55.50	34.80	55.00	55.30
TheoremQA	40.36	30.25	42.38	33.80	42.84	42.77	33.33	42.90	42.77
Llama3.3-70B									
GPQA	47.73	41.41	46.97	44.70	48.74	48.74	45.96	48.74	47.73
MATH500	75.50	58.40	75.60	67.70	76.40	76.40	68.70	76.60	77.10
TheoremQA	65.06	50.33	66.00	59.64	69.01	69.21	59.24	69.55	68.88
AIME25	10.00	0.00	10.00	13.33	13.33	13.33	10.00	13.33	13.33
AIME26	16.67	3.33	16.67	8.33	18.33	18.33	11.67	18.33	20.00
Qwen3-32B									
GPQA	59.09	37.37	59.09	44.44	62.37	62.62	48.48	62.37	64.14
MATH500	86.20	57.40	83.40	74.50	86.40	86.30	72.70	86.60	86.10
TheoremQA	77.51	55.02	79.38	71.49	81.93	81.60	70.89	81.93	81.53
AIME25	80.00	23.33	80.00	58.33	80.00	80.00	55.00	80.00	80.00
AIME26	85.00	50.00	83.33	59.17	87.50	87.50	62.50	87.50	89.17
GPT-OSS-120B: High-Difficulty Evaluation									
AIME25	81.67	16.67	80.00	55.00	85.00	85.00	51.67	85.00	80.00
AIME26	93.33	26.67	90.00	61.67	91.67	90.00	55.00	90.00	85.00
BRUMO25	78.33	23.33	73.33	46.67	75.00	75.00	43.34	75.00	61.67
HMMT25	61.67	16.67	63.33	25.00	65.00	65.00	21.66	65.00	63.33

Table 9 reports Pure ASC and ASC-weighted Vote together with each offset-score variant at $\eta = 0.5$ and $\eta = 1$, isolating direct semantic concentration, answer support, the log-offset construction, answer-frequency normalization, and per-path normalization.

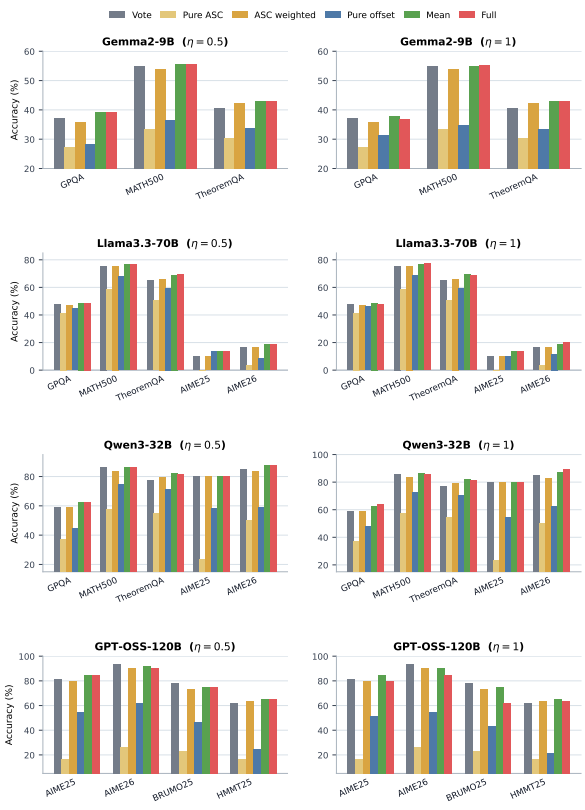


Figure 7: **Cell-level score-component ablation.** Grouped bars report absolute accuracy for Vote, Pure ASC, ASC-weighted Vote, cumulative offset without answer-frequency normalization (Pure offset), the path-normalized Mean variant, and the Full score. Pure ASC averages trajectory concentration within each answer, whereas ASC-weighted Vote sums the same values and therefore retains answer support. Each row compares the same model family at $\eta = 0.5$ and $\eta = 1$; both ASC baselines are independent of η and are repeated as common references.

Figure 7 presents the same variants as grouped bars. Pure ASC underperforms Vote in all 17 cells by 29.37 percentage points on average, showing that semantic concentration without answer support is not a reliable selector. Multiplying the same statistic by answer support produces ASC-weighted Vote, which closely tracks Vote: it is 0.64 percentage points lower on average, improving in five cells, matching Vote in four, and declining in eight. Thus direct semantic weighting is competitive but does not yield a consistent gain. At $\eta = 0.5$, Full improves over ASC-weighted Vote in 15 cells and matches it in two, with an average gain of 2.19 percentage points, isolating the additional value of the relative log-offset construction. Pure cumulative offset, in turn, degrades sharply when its answer-frequency divisor is removed, confirming that the offset must be combined with answer support. Mean and Full isolate effective path length: Mean divides each path's cumulative offset by its total positional weight, whereas Full retains the unnormalized accumulation. At $\eta = 0.5$, they are identical in 11 cells and differ by at most 1.67 percentage points. Their close performance shows that the gain is driven primarily by which semantic states a path visits and the magnitudes of their offsets, rather than by path length or the number of accumulated steps. The larger Mean–Full differences at $\eta = 1$ further indicate that length sensitivity becomes more pronounced when later steps are not discounted. Together, the ablation attributes the improvement to the offset representation and answer-frequency normalization rather than to generic semantic weighting or a preference for longer or shorter trajectories.

Stage 2 semantic containment. We next remove Stage 2 cross-index containment merging, recompute occurrence counts on the Stage 1 same-index overlap-only clusters, and leave all downstream scoring operations unchanged. Table 10 reports the strict accuracy of each model–dataset cell.

Table 10: **Stage 2 semantic-containment ablation by model and dataset.** Accuracy (%) is averaged over repeated runs. Stage 1 clusters same-index steps by overlap (O); the full pipeline additionally merges clusters across step indices by overlap or containment (O/C).

	$\eta = 0.5$			$\eta = 1$		
Dataset	O	O/C	Δ	O	O/C	Δ
			Gemma2-9B			
GPQA	37.87	39.13	+1.26	37.62	36.86	-0.76
MATH500	55.30	55.50	+0.20	54.80	55.30	+0.50
TheoremQA	43.44	42.77	-0.67	43.37	42.77	-0.60
			Llama3.3-70B			
GPQA	47.98	48.74	+0.76	48.24	47.73	-0.51
MATH500	76.40	76.40	+0.00	77.20	77.10	-0.10
TheoremQA	68.81	69.21	+0.40	68.08	68.88	+0.80
AIME25	16.66	13.33	-3.33	13.33	13.33	+0.00
AIME26	18.33	18.33	+0.00	18.33	20.00	+1.67
			Qwen3-32B			
GPQA	62.12	62.62	+0.51	62.88	64.14	+1.26
MATH500	87.00	86.30	-0.70	86.60	86.10	-0.50
TheoremQA	81.80	81.60	-0.20	81.26	81.53	+0.27
AIME25	80.00	80.00	+0.00	76.67	80.00	+3.33
AIME26	90.84	87.50	-3.33	90.84	89.17	-1.67
			GPT-OSS-120B: High-Difficulty Evaluation			
AIME25	86.67	85.00	-1.67	75.00	80.00	+5.00
AIME26	90.00	90.00	+0.00	80.00	85.00	+5.00
BRUMO25	61.67	75.00	+13.33	60.00	61.67	+1.67
HMMT25	61.66	65.00	+3.33	50.00	63.33	+13.33

Table 10 reports O and O/C accuracies and their differences at both positional discounts, showing how containment merging interacts with the model, benchmark, and η.

Figure 8 compares the four absolute accuracies in grouped bars. Containment merging is not a uniform additive correction: its effect is small in many cells but becomes substantial on several high-difficulty GPT-OSS settings, and changing η can alter both its magnitude and direction.

Ablation of the coefficient estimator. Let R denote the later-step repeated-support rate and let $C = \exp(-D)$ denote the second-half versus first-half semantic-centrality factor in Eqs. (73)–(77). We compare the estimated coefficient $\widehat{\eta} = RC$ with an oracle coefficient and the unweighted $\eta = 1$ score.

Table 11 reports the estimated coefficient, the oracle coefficient, and the corresponding estimated, oracle, and $\eta = 1$ accuracies for every model–dataset cell.

Figure 9 compares the estimated, oracle, and no-discount accuracies directly. The estimated coefficient is close to the oracle in several cells, whereas $\eta = 1$ can be materially worse on the GPT-OSS high-difficulty tasks.

Table 12 evaluates calibration-subset sensitivity by independently resampling the unlabeled questions 20 times within every model–dataset cell and evaluating the resulting coefficient in each trial. The resampled coefficient remains concentrated around the original estimate: across all 17 cells, the absolute difference between the original estimate and the resampling mean is at most 0.0550, and the largest coefficient standard deviation is 0.0933. The corresponding test-accuracy standard deviation is at most 0.8567 percentage points and is exactly zero in seven cells. Thus both the estimated coefficient and the resulting prediction accuracy are stable to the particular unlabeled calibration questions selected.

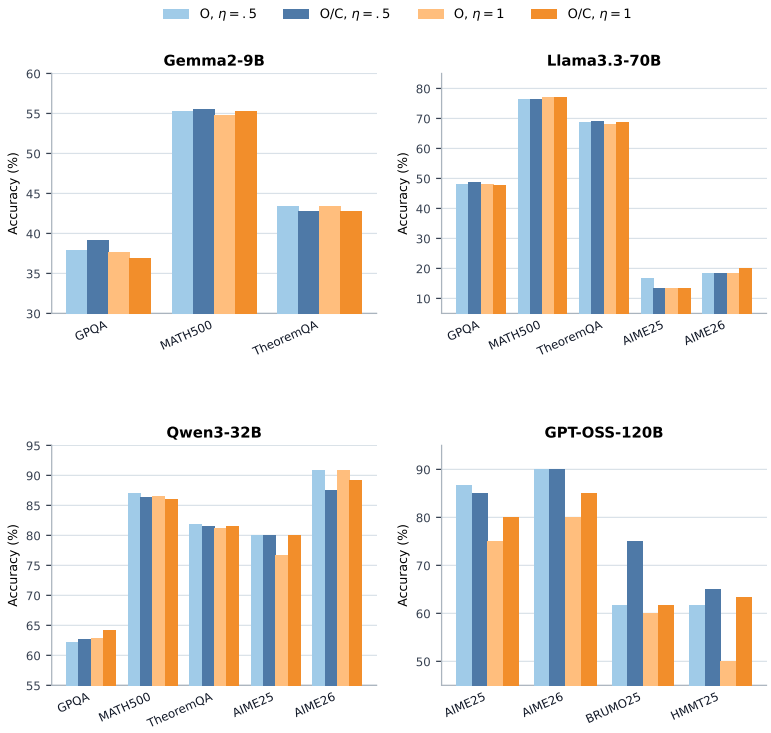


Figure 8: **Cell-level effect of Stage 2 containment merging.** Grouped bars compare same-index overlap clustering (O) with the complete cross-index overlap-containment pipeline (O/C) at $\eta = 0.5$ and $\eta = 1$.

Table 11: **Model–dataset ablation of the positional discount.** For every benchmark, η is estimated from a fixed-seed unlabeled subset containing $\lceil 0.3|Q| \rceil$ questions.

	Estimated		Oracle		No discount
Dataset	η	Accuracy	η^*	Accuracy	$\eta = 1$
Gemma2-9B					
GPQA	.4	39.39	.1	39.90	36.86
MATH500	.5	55.50	.9	55.70	55.30
TheoremQA	.2	42.97	.6	43.51	42.77
Llama3.3-70B					
GPQA	.5	48.74	.6–.8	48.99	47.73
MATH500	.5	76.40	1.0	77.10	77.10
TheoremQA	.5	69.21	.5	69.21	68.88
AIME25	.2	13.33	All	13.33	13.33
AIME26	.3	18.33	.1–.2,.9–1.0	20.00	20.00
Qwen3-32B					
GPQA	.3	62.37	1.0	64.14	64.14
MATH500	.2	86.50	.3	86.60	86.10
TheoremQA	.3	81.73	.4	81.80	81.53
AIME25	.4	80.00	.2–1.0	80.00	80.00
AIME26	.3	87.50	.1	90.83	89.17
GPT-OSS-120B: High-Difficulty Evaluation					
AIME25	.1	85.00	.1,.3–.7	85.00	80.00
AIME26	.1	95.00	.1	95.00	85.00
BRUMO25	.3	75.00	.1–.6	75.00	61.67
HMMT25	.3	66.67	.3–.4	66.67	63.33

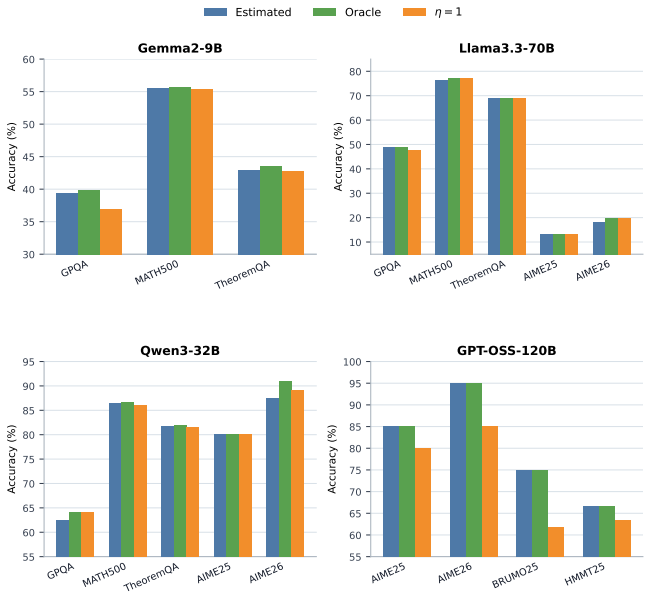


Figure 9: **Cell-level effect of the positional discount.** Grouped bars report the absolute accuracy obtained with the estimated coefficient, the oracle coefficient, and no positional discount ($\eta = 1$).

Table 12: **Stability of the unlabeled coefficient estimator under calibration-question resampling.** For each model–dataset cell, the fixed-size unlabeled calibration subset is resampled 20 times and the resulting coefficient is evaluated normally. The final column reports the observed test accuracy in percentage points. Values are means ± sample standard deviations across resamples.

Model	Dataset	Original η	η mean ± SD	Accuracy mean ± SD
Gemma2-9B	GPQA	0.4	0.3900 ± 0.0718	39.6678 ± 0.3368
	MATH500	0.5	0.5000 ± 0.0795	55.4700 ± 0.0470
	TheoremQA	0.2	0.1700 ± 0.0865	42.3274 ± 0.5623
Llama3.3-70B	GPQA	0.5	0.4850 ± 0.0671	48.7775 ± 0.0916
	MATH500	0.5	0.5050 ± 0.0759	76.4050 ± 0.0759
	TheoremQA	0.5	0.5050 ± 0.0826	69.1163 ± 0.1310
	AIME25	0.2	0.1900 ± 0.0788	13.3300 ± 0.0000
	AIME26	0.3	0.2800 ± 0.0894	19.1650 ± 0.8567
Qwen3-32B	GPQA	0.3	0.2850 ± 0.0745	62.0291 ± 0.7202
	MATH500	0.2	0.2350 ± 0.0671	86.4550 ± 0.2964
	TheoremQA	0.3	0.2800 ± 0.0768	81.7708 ± 0.0342
	AIME25	0.4	0.4150 ± 0.0875	80.0000 ± 0.0000
	AIME26	0.3	0.3150 ± 0.0933	87.5000 ± 0.0000
GPT-OSS-120B	AIME25	0.1	0.1500 ± 0.0513	85.0000 ± 0.0000
	AIME26	0.1	0.1550 ± 0.0510	95.0000 ± 0.0000
	BRUMO25	0.3	0.3000 ± 0.0918	75.0000 ± 0.0000
	HMMT25	0.3	0.2950 ± 0.0759	66.6700 ± 0.0000

G Full Inference-Time Scaling Results

To study finite-sample behavior, we compare SELFOCUS with Vote on identical trajectory prefixes.

Table 13: **Inference-time scaling.** Paired Vote / SELFOCUS accuracy (%).

Dataset	$N = 1$	$N = 2$	$N = 4$	$N = 8$	$N = 16$	$N = 32$	$N = 64$
Gemma2-9B							
GPQA	36.87/36.87	35.73/32.07	30.81/34.85	36.11/38.64	35.86/38.38	37.63/40.40	37.12/39.39
MATH500	46.40/46.40	47.20/43.00	47.30/49.10	51.20/50.80	53.60/53.40	54.60/54.60	54.70/55.50
TheoremQA	40.36/40.36	40.13/36.08	35.94/39.69	39.49/42.24	42.03/44.24	42.24/44.85	40.36/42.97
Llama3.3-70B							
GPQA	44.44/44.44	45.45/42.17	43.69/45.20	46.72/48.23	47.47/48.48	47.22/48.99	47.73/48.74
MATH500	70.30/70.30	71.10/69.90	72.20/72.30	74.10/73.90	75.30/74.80	75.50/74.90	75.50/76.40
TheoremQA	62.12/62.12	62.42/58.97	57.56/63.72	62.25/67.74	63.59/67.60	64.79/68.74	65.06/69.21
AIME25	6.67/6.67	8.33/6.67	6.67/8.33	11.67/11.67	8.33/10.00	10.00/13.33	10.00/13.33
AIME26	8.33/8.33	8.33/6.67	11.67/11.67	15.00/18.33	16.67/16.67	16.67/18.33	16.67/18.33
Qwen3-32B							
GPQA	53.79/53.79	54.92/50.76	53.28/58.33	55.81/60.61	60.10/62.12	59.85/61.36	59.09/62.37
MATH500	79.00/79.00	79.75/77.40	82.50/81.20	83.80/83.00	84.60/84.90	85.60/85.20	86.20/86.50
TheoremQA	72.76/72.76	73.49/68.27	68.14/74.50	71.55/77.38	75.17/79.79	76.51/80.92	77.51/81.73
AIME25	75.00/75.00	71.67/70.00	80.00/80.00	80.00/80.00	80.00/83.33	80.00/78.33	80.00/80.00
AIME26	76.67/76.67	75.83/76.67	81.67/83.33	83.33/83.33	83.33/81.67	85.00/83.33	85.00/87.50
GPT-OSS-120B: High-Difficulty Evaluation							
AIME25	70.00/70.00	72.50/65.00	73.33/75.00	81.67/81.67	83.33/78.33	83.33/83.33	81.67/85.00
AIME26	71.67/71.67	74.17/78.33	78.33/83.33	83.33/86.67	91.67/90.00	91.67/91.67	93.33/95.00
BRUMO25	63.33/63.33	62.50/56.67	61.67/66.67	76.67/75.00	75.00/75.00	76.67/76.67	78.33/75.00
HMMT25	45.00/45.00	43.33/43.33	46.67/50.00	53.33/50.00	56.67/58.33	58.33/56.67	61.67/66.67

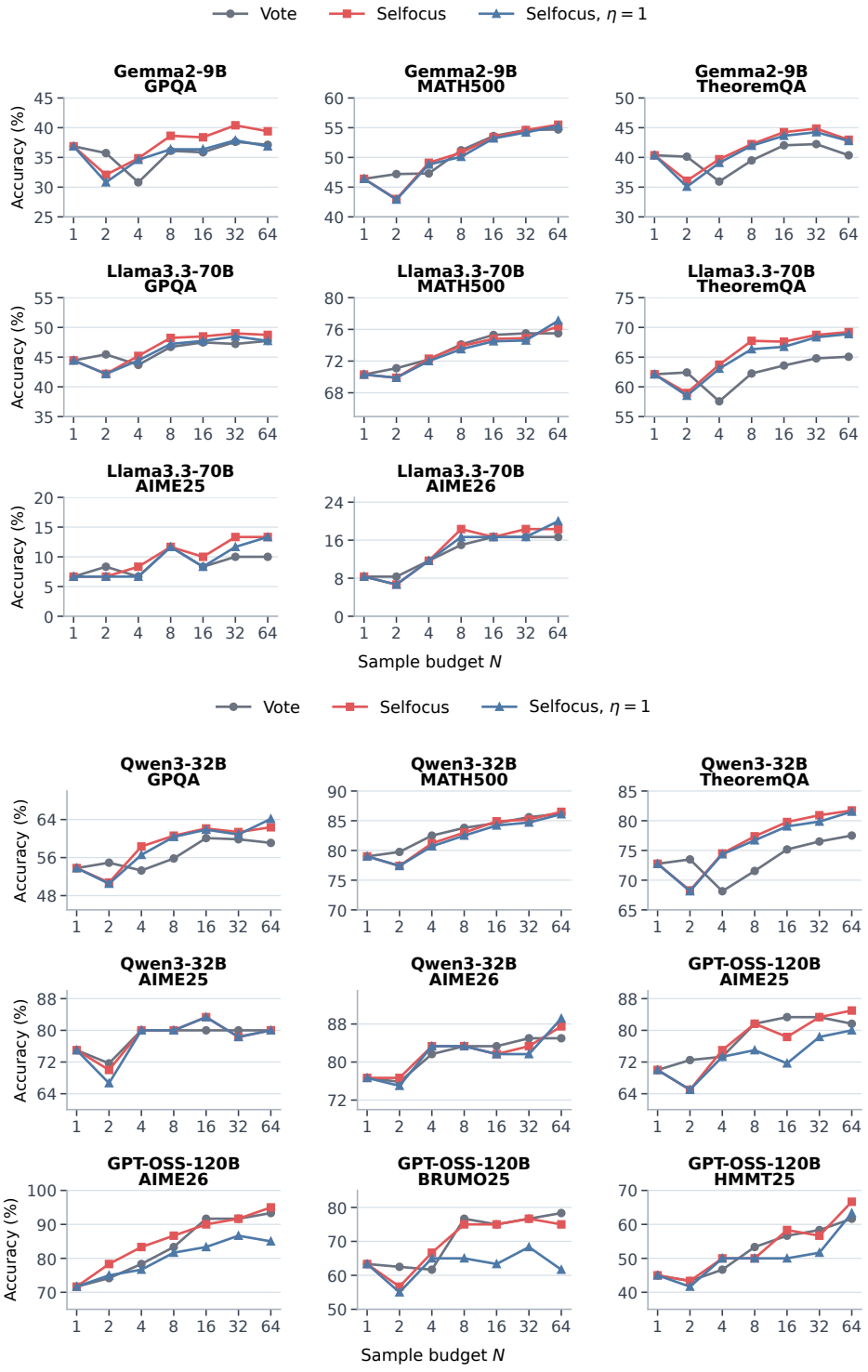


Figure 10: **Scaling with the number of sampled paths for every model–dataset cell.** Each subplot reports paired Vote, SELFOCUS, and unweighted SELFOCUS ($\eta = 1$) accuracy on the same trace prefixes. Repeated evaluations share one $\eta \in \{0.1, \ldots, 1.0\}$ within each model–dataset–budget cell. The upper panel shows Gemma2-9B and Llama3.3-70B; the lower panel shows Qwen3-32B and GPT-OSS-120B.

Figure 10 shows the finite-sample tradeoff across benchmarks. At small budgets, sparse state counts can make the offset unstable and occasionally underperform Vote. Larger budgets reduce this noise, but the trends remain benchmark-dependent: the TheoremQA curves separate more clearly at larger N, the AIME curves move in discrete steps because each benchmark has only 30 questions, and GPT-OSS BRUMO25 does not follow the improving endpoint pattern of GPT-OSS AIME25, AIME26, and HMMT25. This behavior is consistent with the $O_p(N^{-1/2})$ fixed-partition sampling term in Eq. 68: increasing N reduces occurrence-estimation noise but does not by itself remove the reward-conditioned structural residual or the clustering-discrepancy term.

Extended Related Work

Entropy signals in generated reasoning sequences. Recent studies suggest that an autoregressive model's generation distribution is not merely a decoding mechanism, but also an observable of its reasoning process. Across models and tasks, correct or stable trajectories are often associated with more concentrated token distributions, whereas uncertain transitions and failing segments tend to exhibit higher entropy or localized drops in confidence. Although the precise statistic varies across methods, these findings support a common premise: entropy and related probability signals can provide label-free information about reasoning quality (Kang et al., 2025; Fu et al., 2025; Leang et al., 2026; Agarwal et al., 2025). Most work exploits this signal to reduce inference cost. CISC uses confidence-weighted voting to reduce the required number of samples (Taubenfeld et al., 2025); RPC prunes low-probability reasoning paths (Zhou et al., 2025); Think Just Enough, DEER, and DeepConf-Online use entropy or confidence to terminate reasoning early (Sharma & Chopra, 2025; Yang et al., 2026; Fu et al., 2025); and Think Twice Before You Write concentrates branching at high-entropy positions and terminates rollouts using sequence-level entropy (He et al., 2026).

A smaller body of work uses this information primarily to improve final-answer accuracy through post-hoc selection. Self-Certainty measures, at each decoding position, the divergence between the full next-token distribution and a uniform distribution, and aggregates these token-level values into a response-level score for Best-of-N selection (Kang et al., 2025). DeepConf-Offline uses local token-group confidence to filter and reweight completed reasoning traces before answer aggregation (Fu et al., 2025). PiCSAR separately computes the likelihood of generating a reasoning chain and the conditional likelihood of its extracted answer given that chain, then ranks candidates by the sum of these two log-likelihoods (Leang et al., 2026). These three methods are our principal accuracy-oriented baselines because they improve post-generation selection without additional model training. They share our selection objective but use probability information computed within individual traces. Despite their different scoring rules, these methods reduce each completed CoT to a scalar computed from probability statistics internal to that trace. Such local scores may remain high after a model enters a familiar but incorrect derivation, and they do not characterize how independently sampled CoTs relate to one another after multiple trajectories collapse to the same answer. SelFocus instead uses cross-trajectory semantic recurrence to estimate this answer-level structural effect.

Chain-of-thought and inference-time scaling. Chain-of-thought prompting exposes intermediate reasoning states in language models (Wei et al., 2022). Self-consistency improves on greedy decoding by sampling multiple reasoning paths and selecting the most frequent answer (Wang et al., 2023). Universal Self-Consistency concatenates the sampled responses and asks an LLM to select the most consistent candidate, extending self-consistency to free-form outputs without task-specific answer extraction (Chen et al., 2023b). RPC provides a theoretical confidence-estimation framework for sampling-based scaling: it attributes self-consistency's limitation to finite-sample estimation error, identifies modeling error in perplexity-based selection, and combines the two signals with low-probability path pruning (Zhou et al., 2025). Our analysis is complementary. Rather than treating answer selection as confidence estimation, we ask what the answer marginal itself represents under a KL-regularized trajectory policy and isolate reference-policy path volume as a structural source of voting bias.

KL-regularized policy theory. KL-regularized reinforcement learning and preference optimization provide the theoretical starting point for our analysis. RLHF and related preference-training methods optimize sequence-level reward while regularizing the learned policy toward a reference model (Ouyang et al., 2022). The corresponding optimum has a Gibbs form: reward exponentially

tilts the reference distribution over complete responses. DPO makes the same structure explicit by relating the reward of a complete response to its policy–reference log-ratio (Rafailov et al., 2023). RL objectives have subsequently become central to mathematical and chain-of-thought reasoning models. DeepSeekMath applies reinforcement learning to improve mathematical reasoning, while DeepSeek-R1 demonstrates that large-scale RL can elicit extended reasoning behavior and substantially improve performance on difficult reasoning tasks (Shao et al., 2024; DeepSeek-AI, 2025).

Building on this sequence-level Gibbs formulation, we decompose each generated response into a sequence of reasoning actions followed by a final answer, and marginalize all trajectories that lead to the same answer. This yields an answer-level distribution in which answer frequency depends not only on outcome reward but also on the reference-policy mass of its supporting trajectories. Our theoretical extension identifies this path-volume term as a possible source of self-consistency bias. Separately, SELFOCUS estimates a post-RL occurrence functional from cross-trajectory semantic recurrence. Only under the explicit reward-conditioned residual assumption A2 does this functional admit a partial path-volume interpretation; the benchmark results establish improved re-ranking, not the mechanism itself.

Verification and path-aware aggregation. Verifier-based approaches score final answers or intermediate steps (Cobbe et al., 2021; Lightman et al., 2024; Weng et al., 2023). AoR explicitly targets cases in which the correct answer is a minority: it hierarchically evaluates reasoning chains and dynamically allocates additional samples before selecting an answer (Yin et al., 2024). Semantic Self-Consistency embeds each complete reasoning path and weights answer aggregation by centroid proximity or pairwise cosine consensus (Knappe et al., 2025). Reasoning-aware self-consistency combines path quality with answer consistency for efficient sampling (Wan et al., 2025). Slim-SC uses thought-level inter-chain similarity to prune redundant trajectories during generation (Hong et al., 2025), whereas RISC learns an answer ranker from frequency, semantic centrality, and reasoning-trace consistency features (Marina et al., 2026). At training time, CoVo uses the consistency and volatility of intermediate states across trajectories as an intrinsic reward for self-rewarding RL (Zhang et al., 2025). SELFOCUS uses this structural information for post-generation accuracy improvement without training a verifier, ranker, or policy. Repeated semantic step occurrences provide a relative free-energy offset for answer re-ranking, with the generator and semantic models kept frozen and no token log probabilities required. Our objective is more accurate selection from sampled solutions, rather than shorter generations.

DISCUSSION

What CoT-level structure adds beyond token confidence. Token-derived confidence measures how decisively the model generates a trace, whereas semantic occupancy measures recurrence across independently sampled CoTs. Paraphrases of one derivation can have different token likelihoods yet share a semantic mode; incompatible derivations can each be locally confident while forming a diverse answer-conditioned path family. CoT-level aggregation therefore distinguishes reasoning consensus from endpoint agreement. Semantic concentration does not universally imply correctness: a wrong answer may be supported by a coherent, recurrent misconception, causing SELFOCUS to fail.

What offset estimates. For the fixed reference partition used in the analysis, the population offset $o(c)$ equals the relative occurrence-space free energy $\beta(F_c - F_{c^*})$ in Eq. 38. Its empirical plug-in estimate has sampling error controlled by Eq. 50. The implemented O/C partition is sample-dependent, and Eq. 41 separately records the resulting answer-level discrepancy.

Why a reference model matters. The strict decomposition requires Q_y, the reference-policy mass of paths ending in answer y, rather than the post-RL vote probability. Under the fixed reference partition, the empirical mean offset consistently estimates G_y, an answer-conditioned occurrence functional. The implemented statistic additionally differs by $\Delta^{\text{clust}}_{N,y}$, which is assumed bounded but not universally vanishing. Post-RL samples therefore do not identify Q_y without further structure. A2 expresses the remaining population gap as a reward-conditioned component plus an answer-specific residual. Equation (56) bounds the resulting partial estimate, and Eq. (68) separates the nonvanishing structural residual, the $N^{-1/2}$ fixed-partition sampling term, and the clustering term κ_N. Semantic concentration alone therefore does not identify reference answer mass.

The Gibbs analysis describes the idealized outcome-only KL-regularized objective rather than every deployed RL-trained model. Increasing the number of sampled trajectories reduces finite-sample estimation noise but does not remove representation, clustering, or structural approximation error. Accordingly, the benchmark evidence supports the empirical statement that cross-trajectory semantic recurrence improves answer re-ranking. Path-volume correction is one theoretical explanation when A2 holds, not a mechanism identified by those experiments.

Relation to self-consistency. Self-consistency works when correct reasoning also dominates marginal mass. The failure mode appears when the answer projection merges many independent errors into one option or separates equivalent correct strings into several buckets. SELFOCUS combines semantic recurrence with answer frequency to re-rank the sampled candidates.

Why early weighting can help. An early semantic deviation can redirect all subsequent steps, while late algebraic elaboration may add length without changing the answer basin. This motivates $\eta < 1$ as a depth-weighted composite-likelihood aggregation rule applied to the offsets in Eq. 38.