THE MONTY GAP: SELECTION EVIDENCE AND ITS LIMITS

Anonymous authors
Paper under double-blind review

ABSTRACT

A verifier's choice of which answer to reject can carry evidence beyond the rejection itself. We audit whether this selection evidence improves model decisions: no tested prompted critic establishes a positive subject-cluster-robust accuracy gain over re-prompting, and a separate confidence study finds no reliable decision-cost benefit over calibrated controls. To interpret these results, we define the *Monty gap*, the expected posterior information discarded by masking rejected answers. We prove that every conclusive verifier that always spares the model's commitment at a positive fixed budget is non-ignorable. This guarantees a posterior discrepancy, not a decision gain. Under uniform selection among wrong alternatives, switching has a simple threshold: the surviving rival's prior probability relative to the commitment's must exceed $(N-1-k)/(N-1)$. We characterize when masking is exact and show how intermediate discrepancies can disappear over a fixed budget. Scripted oracles establish the mechanism on model output distributions; prompted critics violate its assumptions, and the only positive four-option pairing is concentrated in one subject block. The resulting contribution is an audit of verifier protocols and their limits: selection evidence can exist without being reliably usable by the decoder.

1 INTRODUCTION

A verifier tells a model that one of its candidate answers is wrong. A natural update is to remove that answer and renormalize the remaining probabilities. This operation, which we call *masking*, uses the content of the report: this answer is wrong. There is also a choice to interpret: why did the verifier reject this answer rather than another? That choice may reflect what the verifier knows about the truth. This paper asks when the second message is worth reading.

The Monty Hall problem illustrates the distinction. A host who knows the prize's location deliberately opens an empty door; a host who opens a door by accident can produce the same observation with a different posterior [1; 2]. In model evaluation, the corresponding question is whether the verifier's choice among possible corrections conveys information beyond the correction itself. This depends on how the verifier selects its reports, not just on whether a report is correct.

The distinction becomes concrete when the model already favors one answer. Suppose $p(A) = 0.80$, $p(B) = 0.19$, and the remaining 0.01 is spread equally across the other candidates. A verifier knows the truth and samples uniformly from the wrong candidates other than the model's commitment A. With four candidates, ruling out C leaves A ahead. With a hundred candidates, ruling out all 98 alternatives to A and B leaves B ahead. Masking still prefers A in both cases. Repeatedly sparing B is much more likely when B is correct. After enough eliminations, that survival outweighs the model's initial confidence in A. The exact rule is

$$\frac{p(y)}{p(c)} > \frac{N-1-k}{N-1},$$

where c is the commitment and y the best surviving alternative after k eliminations. The rule requires the stated uniform policy. Merely instructing a critic to spare c does not establish that policy.

The empirical question has a negative answer in the tested settings: no prompted critic establishes a positive subject-cluster-robust accuracy gain over re-prompting. The only positive four-option

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.

pairing is concentrated in one subject block; ten-option critics often violate commitment sparing; and a separate fixed-answer confidence study finds no reliable cost benefit over calibrated controls. These findings motivate an audit of the protocol rather than a claim of a generally useful decoding method.

The theory separates three questions that this audit requires. Does selection change the posterior? Must it change the chosen answer? Can its likelihood be recovered well enough to improve measured decisions? Every conclusive, commitment-sparing kernel with a positive fixed budget fails the exact-masking condition. A uniform such kernel yields the threshold above, but neither non-ignorability nor the threshold guarantees an empirical benefit. A verifier that already knows the answer could simply report it if the interface permits; our oracle experiments hold an elimination interface fixed to isolate the information in selection. They do not establish that this interface is preferable.

Related work and scope of the contribution. Bayesian conditioning on the observation protocol is classical [8; 9]. Masking is exact under the coarsening-at-random (CAR) condition [7; 23]; Proposition 1 is its specialization to elimination reports. Gill and Grünwald's CAR polytope includes deterministic partition mechanisms. Our block classification specializes that structure to ordered, fixed-size elimination sequences: it yields the divisibility condition in Theorem 1, with counts and dimension deferred to Supplement J. It is not an alternative general characterization of CAR. The information identity is a support-restricted KL decomposition [14], and the switching rule is a likelihood-ratio calculation. We use these established tools to connect a verifier's selection rule to a measurable discrepancy and a decision threshold.

Host-dependent posteriors and odds-ratio updates also appear in the Monty Hall and restricted-choice literature [2; 28; 29]. Rosenhouse [49], including his treatment of progressive Monty, discusses extensions beyond three doors; we do not claim that multiple eliminations or non-uniform beliefs are new. The focus here is the classification of elimination protocols and the empirical separation of report content, selection, and decision effects. Process-of-elimination prompting [3; 4; 41] and conformal pruning followed by re-prompting [37] provide relevant inference settings. Work on self-correction [38] asks when feedback changes answers successfully; our analysis asks what additional evidence the choice of feedback supplies. Code selection through execution agreement [42; 43; 44] is especially useful for distinguishing the content of a failure report from its selection.

2 Setting

There are $N \geq 3$ candidates $\mathcal{Y} = \{1, \ldots, N\}$, one correct answer Y, and a full-support model belief p. The commitment c maximizes p, with a fixed tie-breaking rule. An oracle knows Y and reveals k distinct wrong candidates, where $1 \leq k \leq N-2$. Its conditional policy may depend on Y, p, c, and the previous reveals. Holding p and c fixed, write $L_\pi(r \mid y)$ for the likelihood of the complete sequence $r = (r_1, \ldots, r_k)$ and $S(r) = \mathcal{Y} \setminus \{r_1, \ldots, r_k\}$ for the survivors. The two posteriors are

$$m_r(y) = \frac{p(y)\mathbf{1}\{y \in S(r)\}}{p(S(r))}, \qquad q_r(y) = \frac{p(y)L_\pi(r \mid y)}{\sum_z p(z)L_\pi(r \mid z)}. \tag{1}$$

Here m masks the revealed candidates and q uses their likelihood. Sequence likelihoods are products of conditional round-wise likelihoods, evaluated at the observed history.

The information results use the *subjective joint* $P_p(Y = y, R = r) = p(y)L_\pi(r \mid y)$: they describe an update coherent with the model's belief. Experimental accuracy uses the dataset's answer key instead. We denote the measured accuracy difference in percentage points by Δ_{acc}; this is distinct from a subjective decision gain in probability units.

Three policies organize the main analysis. *Uniform-among-wrong* samples uniformly from unrevealed wrong candidates. *Avoid-commit* additionally excludes c, and is uniform over the remaining legal choices. An *eager* oracle reveals the highest-ranked wrong candidate. Rankings are held fixed during an episode. A reverse-rank (*spiteful*) control instead reveals the lowest-ranked wrong candidate; it is retained in the technical supplement. Prompted critics can violate both the selection rule and conclusiveness; their decoder and error-handling conventions are specified in Appendix B.

3 THEORY

3.1 EXACT MASKING AND THE MONTY GAP

Proposition 1 (Exactness of masking). *For full-support p and a sequence r of positive probability, $q_r = m_r$ if and only if $L_\pi(r \mid y)$ is constant over $y \in S(r)$.*

The factor $p(y) > 0$ cancels in equation (1). Thus masking is exact precisely when the report is equally likely under every surviving truth. We call a kernel satisfying this condition at every possible report *ignorable*.

Theorem 1 (Ignorable elimination protocols). *For one reveal, the only ignorable kernel is uniform among wrong answers. For a fixed budget k, a deterministic reveal map is ignorable if and only if it is a* block protocol*: partition $\mathcal{Y}$ into blocks of size $N - k$, and let every truth in a block reveal its complement in the same fixed order. Such maps exist if and only if $N - k$ divides N.*

For one reveal, write the common likelihood in column r as κ_r. The row sums $\sum_{r \neq y} \kappa_r = 1$ imply that all κ_r are equal. For a deterministic map F, the posterior restricts p to the fiber $F^{-1}(r)$, while masking restricts it to $S(r)$. Full support makes them equal exactly when these sets agree, giving the block partition. Appendix A expands this argument; Supplement J gives the protocol count, kernel-polytope dimension, and finite-case certificates. If the policy reads p, the condition applies separately to each induced map or kernel.

A small example shows why a budget matters. At $N = 4, k = 2$, let truths 1 and 2 produce $(3, 4)$, and truths 3 and 4 produce $(1, 2)$. After observing 3, the policy already identifies the block $\{1, 2\}$, while masking retains $\{1, 2, 4\}$. After observing 4, both retain $\{1, 2\}$. Later reveals can therefore remove an earlier discrepancy. The gap need not accumulate with the number of reveals (Supplement J).

Theorem 2 (Information decomposition). *Under P_p, with natural logarithms,*

$$I_p(Y; R) = \underbrace{\mathbb{E}_R[-\log p(S(R))]}_{\textit{elimination information}} + \underbrace{\mathbb{E}_R[D_{\mathrm{KL}}(q_R \| m_R)]}_{G\textit{: Monty gap}}. \tag{2}$$

The gap is nonnegative and vanishes exactly for ignorable kernels.

To see the decomposition, split $\log(q_r/p) = \log(q_r/m_r) + \log(m_r/p)$. On the posterior's support the second term is the constant $-\log p(S(r))$. Average first over $Y \mid R = r$, then over R.

Corollary 1 (Excess log-loss). *The excess expected log-loss of masking under P_p is exactly the gap:*

$$\mathbb{E}_{P_p}[-\log m_R(Y)] - \mathbb{E}_{P_p}[-\log q_R(Y)] = G.$$

The first term in equation (2) accounts for learning that the revealed answers are wrong. The second accounts for learning that the verifier selected those answers. Neither the identity nor its nonnegativity promises an improvement against an answer-generating distribution different from p.

Theorem 3 (Commitment sparing prevents exact masking). *Fix a commitment c, full-support p, and $1 \leq k \leq N - 2$. If a conclusive kernel always reports k distinct candidates other than c, it is non-ignorable and has $G > 0$ under P_p.*

Proof. If it were ignorable, write $L(r \mid y) = \kappa_r \mathbf{1}\{y \in S(r)\}$. Every possible report spares c, so the row for c gives $\sum_r \kappa_r = 1$. Summing all N rows instead gives $(N-k)\sum_r \kappa_r = N$, a contradiction for $k > 0$. Theorem 2 then gives $G > 0$. □

The guarantee concerns a fixed conclusive kernel, not a prompt requesting one. It does not imply a discrepancy on every realized report, identify the kernel from logs, or imply a different decision. Nor does it extend unchanged to fallible verdicts or variable stopping budgets.

3.2 INFORMATION AND DECISIONS

A changed posterior need not change the selected answer. Conversely, a model can retain its original answer despite many correct eliminations. We state the decision rule first, then these two limiting cases.

Proposition 2 (Switching under avoid-commit). *Under the uniform avoid-commit policy, define* $B = (N-1)/(N-1-k)$. *For each surviving* $y \neq c$,

$$\frac{q_r(y)}{q_r(c)} = B\frac{p(y)}{p(c)}. \tag{3}$$

With ties resolved in favor of c, *the decoder leaves* c *if and only if the best surviving alternative satisfies* $p(y)/p(c) > 1/B$.

If c is true, each reveal is drawn from the $N-1$ alternatives; if $y \neq c$ is true, y must also be spared. The ratio of the two ordered-sequence likelihoods telescopes to B. All surviving alternatives receive this same multiplier, so their relative order is unchanged.

Corollary 2 (Temperature correction). *For* $p_T(y) \propto p(y)^{1/T}$, *applying Proposition 2 to* p_T *is equivalent to comparing the raw ratio* $p(y)/p(c)$ *with*

$$\tau_T(N,k) = \left(\frac{N-1-k}{N-1}\right)^T. \tag{4}$$

This is an algebraic identity; whether temperature scaling adequately calibrates a model is empirical. A fitted $T > 1$ lowers the raw-ratio threshold, and $T < 1$ raises it. Neither corrects an incorrectly assumed reveal policy.

Proposition 3 (Commitment preservation). *If the verifier never reveals* c, *masking retains* c *at every budget under a fixed tie-breaking rule. Its accuracy is the unaided accuracy. Avoid-commit satisfies this hypothesis.*

Together, Theorem 3 and Proposition 3 locate guaranteed selection information in a setting where masking is guaranteed to retain its unaided answer and accuracy, even though its confidence can change.

Theorem 4 (Eager decisions). *Rank candidates by decreasing* p *with fixed tie-breaking. If the truth has rank* j, *eager elimination makes masking and the policy-aware decoder choose the same answer at every budget, correct exactly when* $j \leq k+1$.

An eager oracle's skipped candidate, when there is one, is already the highest-ranked survivor. Selection can thus improve the posterior without improving top-1 accuracy.

Proposition 4 (Subjective accuracy bound). *Let* $a_q(r)$ *and* $a_m(r)$ *maximize* q_r *and* m_r. *The subjective gain*

$$\delta_p = \mathbb{E}_{P_p}[\mathbf{1}\{a_q(R) = Y\} - \mathbf{1}\{a_m(R) = Y\}]$$

satisfies $0 \leq \delta_p \leq \sqrt{2G}$.

This follows from a $2\|q_r - m_r\|_{\mathrm{TV}}$ decision bound and Pinsker's inequality (Appendix A). It does not bound Δ_{acc} on miscalibrated beliefs. The eager example rules out a converse: positive selection information need not buy a different decision.

3.3 The confidence threshold

For the opening example, $p(B)/p(A) = 0.19/0.80$. With four candidates and one elimination, $B = 3/2$ and the posterior still favors A. With a hundred candidates, conditional on B surviving, the first switch occurs at $k = 76$ because $99/(99-k) > 80/19$. After 98 eliminations the posterior on B is $99(0.19)/(0.80 + 99(0.19)) \approx 0.959$. Masking assigns A probability $0.80/0.99 \approx 0.808$ and keeps it.

Keeping $p(A) = 0.80$ but spreading 0.20 uniformly over the other 99 candidates produces a different answer. After 98 avoid-commit eliminations, the policy-aware posterior remains $(0.80, 0.20)$ on the final pair, and keeps A. Masking raises A to approximately 0.9975. The commitment probability alone is therefore insufficient: the distribution over alternatives matters. Supplement A gives both calculations. Figure 1 shows the threshold as a function of the eliminated fraction.

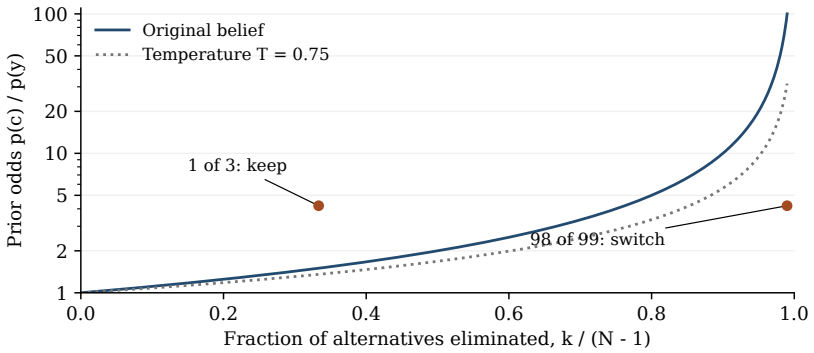

Figure 1: Switching under the uniform avoid-commit policy. The boundary is $p(c)/p(y) = (1 - k/(N-1))^{-1}$; the policy-aware decoder switches below it and keeps the commitment above it. The dotted curve applies the temperature correction at $T = 0.75$. The points use $p(c) = 0.80, p(y) = 0.19$, after one of three alternatives or 98 of 99 alternatives have been eliminated.

4 Experiments

We separate three questions: how large the decision effect is under a known policy, whether prompted critics follow such a policy, and whether a richer report helps through its content rather than its selection. Scripted oracles read the answer key and obey their policy exactly. They provide controlled demonstrations on real belief distributions, not evidence of a deployable improvement or an upper bound over all possible verifiers. Critic studies measure policy adherence and compare decoders on the same observed eliminations.

Unless specified otherwise, intervals are paired 95% percentile bootstrap intervals over items. They resample saved correctness differences without refitting the policy tables, so they condition on the fitted decoders. We also report subject-cluster intervals where available: these resample subject blocks with replacement and compute the item-weighted mean across the sampled blocks. With only ten MMLU or twelve MMLU-Pro clusters, nominal coverage is uncertain and results are sensitive to the observed subject mix; item intervals cannot resolve this limitation. Appendix B specifies the decoder, cross-fitting, and donor-rank resampling control. We display empirical accuracy effects to one decimal place; exact outputs remain in the artifact. These digits identify observed differences, not the precision of generalization to new subjects. The separate technical supplement retains full results and experimental provenance under its original A–N labels.

What the belief measurement means. The main experiments normalize next-token option-letter probabilities. These are model output distributions, not a verified account of internal belief or calibrated correctness. Option position matters: Supplement G records Qwen2.5-7B's marginal $P(B) = 0.341$. Four-permutation averaging mitigates position effects but does not remove this conceptual limitation. BF16 ties also affect rank-based policies (Supplement H). The avoid-commit threshold reads probability ratios, so errors in magnitudes matter even if the leading option is stable. Temperature scaling is a sensitivity check, not validation of the belief object.

4.1 Prompted critics

Four options. A Qwen2.5-14B-Instruct critic sees the question, the student's commitment, and usually the answer key. It is asked to eliminate a wrong alternative while sparing the commitment, but does not see the student's probabilities. The primary student is Llama-3.1-8B on 4,000 MMLU items. Follow-ups change the critic to phi-4, the student to Qwen2.5-7B, the key, or the reveal budget.

The *estimated* decoder fits two reveal-role tables on the opposite data fold: one when the hypothesized truth equals the commitment and one when it does not. The *rank-flattened* decoder averages the non-commitment rank cells within those tables. It removes those rank preferences, but retains fitted truth and commitment rates. It therefore remains informed by the training keys and is not a

Critic × student or control	Δ_{acc}	95% item interval
Qwen-14B × Llama-8B, rank-flattened	$+2.2$	$[+1.0, +3.2]$
same pairing, full estimated table	$+3.3$	$[+1.8, +4.8]$
phi-4 × Llama-8B, rank-flattened	-0.1	$[-0.9, +0.8]$
Qwen-14B × Qwen-7B, rank-flattened	-0.6	$[-1.4, +0.2]$
phi-4 × Qwen-7B, rank-flattened	-0.4	$[-1.0, +0.2]$
Keyless critic, rank-flattened	-0.2	$[-0.7, +0.4]$
Sham key, estimated table	-0.4	$[-1.0, +0.1]$
Sham key, assumed avoid-commit	-2.9	$[-5.7, -0.1]$

Table 1: Prompted critics at $N = 4, k = 1$: decoder minus masking, in percentage points. The sham-key run uses the first 1,000 items, an easier subset, so it is a within-subset control rather than a matched contrast with the full real-key run. Full controls and resampling diagnostics appear in Supplement C.

complete separation of selection from critic competence. Appendix B gives the exact cells, estimator, fallback, and resampling procedures.

The four-option study does not establish a subject-robust benefit. Only one of the four pairings gives a positive rank-flattened point estimate (Table 1). Its approximately $+2.2$ points come entirely from `moral_scenarios`: $+11.6$ on 895 items, versus -0.6 on the other 3,105. Unaided accuracy on that block is 29.1%, against a 25% uniform-choice reference. On 892 of those 895 items the runner-up ratio exceeds $2/3$, the ideal policy's one-reveal threshold. This condition means the two leading probabilities are relatively close; it does not imply a nearly uniform four-option belief. The ten-subject cluster interval includes zero. We treat this as a subject-specific diagnostic, not positive evidence of a general critic benefit. Excluding the block is a post-hoc sensitivity analysis; it cannot be relabeled a prospectively held-out test. A repeated greedy critic pass gives $+1.9$ points on the same item pool.

The subject decomposition also reveals a policy mismatch. Among the 1,093 items outside `moral_scenarios` that exceed the ideal switching threshold, the critic yields -1.7 points, whereas a scripted avoid-commit oracle on the same beliefs yields $+7.5$. When the commitment is correct on these items, the critics tend to strike the least-favored option rather than sample uniformly. The surviving runner-up is then weaker evidence than the ideal rule assumes. Supplement K reports the rank distributions and confidence strata. They are consistent with this explanation, but do not identify a single causal source of the between-pairing differences.

A second-reveal run returned a distinct second strike on only 40.6% of items and failed its registered validity condition. It cannot establish the predicted budget effect. Removing the answer key gives no positive rank-flattened gain; this supports the importance of information supplied by the key in this setting, but the keyed phi-4 null shows that access to the key is not sufficient.

Ten options. On MMLU-Pro, struck options are removed from later prompts. Five registered cells combine three students with three critics: Qwen2.5-14B-Instruct, gemma-2-9b-it, and Llama-3.1-8B-Instruct. Each cell has 2,000 items and budgets 2, 4, and 8. The critic receives the answer key; "log-only" below describes the estimator, not the critic's access to truth. The primary decoder fits a rank-flattened policy from logs without training keys. It uses the student's belief as a latent-truth prior and assumes conclusive reports, so calibration and false eliminations remain sources of misspecification. Re-prompting uses the same eliminations. Its four arms cross retaining versus relabeling the surviving option letters with omitting versus adding an explicit notice that an expert ruled out the rejected options. Table 2 selects the arm with highest evaluation accuracy within each cell and budget. Its intervals condition on that selected arm; they do not evaluate a prospectively fixed arm-selection procedure.

Across the five cells, critics strike the commitment by round eight on 65–91% of items whose initial commitment is wrong; this percentage is conditional on an initially wrong answer, not all items. Thus the prompted protocol does not implement Theorem 3. For Qwen-14B and Llama-Instruct, the estimated decoder differs from masking by at most 0.1 points at budgets 2 and 4, then loses 4.5–8.0 points at budget 8. Gemma follows the first-round instruction more often, yet gives mixed results: -2.1 and -2.5 at budget 4, then $+2.6$ and $+0.8$ at budget 8. Neither positive result clears a

Student × critic	vs. masking	vs. re-prompting
Llama-8B × Qwen-14B	-4.5 $[-6.3, -3.1]$	-6.5 $[-8.7, -4.2]$
Mistral-7B × gemma-9B	$+2.6$ $[-0.9, +6.2]$	$+1.8$ $[-1.3, +4.4]$
Qwen-7B × gemma-9B	$+0.8$ $[-0.8, +2.5]$	-0.2 $[-2.9, +2.0]$
Mistral-7B × Llama-8B-Instruct	-8.0 $[-10.2, -5.4]$	-5.6 $[-8.2, -3.4]$
Qwen-7B × Llama-8B-Instruct	-5.5 $[-7.3, -3.0]$	-2.8 $[-5.1, -0.5]$

Table 2: Log-only rank-flattened decoding at $k = 8$ on MMLU-Pro: accuracy differences in percentage points, with 95% intervals resampling twelve subject categories. Re-prompting is the best of four arms on identical eliminations. No interval against either comparator is wholly positive (Supplement L).

subject-cluster interval. The decoder that assumes the prompted policy loses 6.5–18.6 points across all five cells. Masking itself improves on the prior by 19.1–28.5 points because the commitment is frequently eliminated. These are empirical results for imperfect critics, not instances of the exact eager-oracle theorem. The ten-option study does not repair the validity failure of the earlier four-option run.

What log-only estimation establishes. At four options, constrained EM recovers the keyed commitment-rate estimate closely for the Llama student, but overestimates it for the overconfident Qwen student. Its largest rank-flattened gain is $+2.6$ with a ten-block interval of $[-1.3, +8.4]$. Neither of the two matched re-prompting comparisons is positive by its item interval (Supplement C). These comparisons retain only items where the critic and a scripted oracle struck the same option. This selects compliant, conclusive reveals and overrepresents initially wrong commitments; it is not a comparison on the full critic population. These fits demonstrate a possible estimation procedure under calibration and structural assumptions. They do not establish that an arbitrary policy, or its ignorability, is identifiable from reveal logs alone [22].

4.2 Scripted policies on real beliefs

Four base models (Qwen2.5-7B and 1.5B, Llama-3.1-8B, and Mistral-7B-v0.3) supply option-letter beliefs on MMLU, ARC-Challenge, and a four-option construction of TruthfulQA-MC1. Temperature fitting and evaluation use separate halves. For TruthfulQA, items with fewer than four options are dropped; retained items use the correct answer and the first three wrong answers in stored order, with seeded option shuffling. Thus these results do not cover the native variable-option task.

The policy-aware and masking decisions agree exactly under uniform-among-wrong and eager elimination, as the theory requires. Under avoid-commit at $N = 4, k = 2$, the gains are 6.1–19.2 points on MMLU, 1.7–3.9 on ARC, and 4.7–11.7 on the TruthfulQA construction (Supplement D). The smaller evaluations contain 583 and 342 items respectively; three of their eight avoid-commit McNemar tests have $p > 0.05$. These runs quantify the rule's effect on the measured belief distributions. They do not test whether an actual verifier supplies the likelihood used by the decoder.

Misspecifying that likelihood can reverse the effect. In the scripted grid, assuming avoid-commit against a uniform or eager oracle loses as much as 28.2 points. Assuming uniform-among-wrong is exactly masking. The assumed-eager decoder also agrees with masking in this grid, through either an argmax-preserving update or the zero-likelihood fallback in Algorithm 1. This is a property of these specified decoders, not a general ordering in which less structured assumptions are always safer.

4.3 Checks on the source of information

A fixed test suite with known reference outputs supplies no selection evidence if every live candidate is evaluated on each administered test and test selection depends only on the observed history. All survivors then produce the same observed outputs (Supplement J). With hidden references, shared wrong outputs can eliminate additional programs. The base pool contains 7,857 Qwen2.5-Coder-7B program samples across 164 HumanEval tasks. Evaluation retains tasks containing both correct and wrong candidates: 140 base-model tasks and 79 instruct-model tasks. At ten verifier executions,

richer reports yield $+6.4$ pass@1 points $[+2.9, +10.7]$ and $+3.8$ $[+0.0, +8.9]$ on those respective subsets. These are report-content gains: the separate survival-evidence contrast is $+3.6$ on the base pool and zero on the instruct pool, without a replicated selection benefit (Supplement E). Published-order and reverse-rank oracle sweeps are retained only as controls showing what an answer-aware protocol can encode (Supplements C–D). Their gains are not evidence for a useful critic.

4.4 Calibration and fallible verdicts

Under scripted avoid-commit, the temperature-corrected threshold is closer to the empirical optimum in 54 of 72 dataset–model–budget cells. It improves accuracy over the raw threshold by a median 0.8 points, with the largest changes on underconfident MMLU-Pro beliefs (Supplement D). Under real four-option critics, the same correction changes accuracy by -1.3 to $+0.9$ points. Correcting the belief does not supply a missing policy model.

The conclusive decoder assigns zero weight to a falsely eliminated truth. An explicit error component instead assigns positive likelihood to such a verdict (Algorithm 1). In the reported noise analysis, the resulting decoder is correct on 44.0% of falsely eliminated items, versus zero under hard elimination. This protects against errors under the specified mixture; it does not validate uniform errors as a model of real critic mistakes.

4.5 Confidence and alternative belief measurements

A separate keyless Qwen2.5-14B-Instruct run mechanically enforces eight distinct commitment-sparing choices on 1,200 shared MMLU-Pro questions for each of Mistral-7B and Qwen-7B. Policy fitting, confidence calibration, and evaluation use disjoint groups of 400 questions each. All methods retain the same answer. Answers with calibrated confidence at least 0.8 are accepted. The primary loss charges one for an accepted error and 0.20 for escalation, assuming perfect escalation and excluding critic computation cost. Separate calibrators and soft-elimination, prior, and generic feature controls distinguish selection evidence from calibration and fallible verdicts. Two hundred subject-bootstrap replicates refit the policy and calibrators.

Expressed per 1,000 decisions, selection-aware cost minus calibrated masking is -2.5 $[-7.1, +4.0]$ for Mistral and -0.5 $[-16.3, +12.7]$ for Qwen. Neither pool meets the planned success criterion against all three required controls. The aware and masking arms accept only about 6% of Mistral items and 14% of Qwen items, leaving mean cost dominated by escalation; the Mistral contrast turns on one accepted error. This limited resolution does not establish absence of a benefit. Table 3 reports all calibrated controls. The critic strikes the truth on 42% and 38% of evaluation items respectively, so mechanical commitment sparing does not recover the conclusive setting. The narrower Mistral log-loss improvement over calibrated masking does not beat simpler controls in point loss. Supplement N gives the full design and results.

Student	Calibrated confidence	Cost per 1,000	Accepted	Errors
Mistral-7B	Masking	195.0	25	3
	Selection-aware	192.5	25	2
	Soft elimination	192.0	26	2
	Prior	191.5	27	2
	Generic	194.0	27	3
Qwen-7B	Masking	185.5	54	5
	Selection-aware	185.0	55	5
	Soft elimination	186.0	53	5
	Prior	186.5	52	5
	Generic	180.0	55	3

Table 3: Fixed-answer confidence, 400 evaluation items per student; acceptance threshold 0.8. Errors count accepted wrong answers. Lower cost is better; scaling the mean cost by 1,000 changes units, not uncertainty or sample size. Students share questions. The full-refit intervals for selection-aware differences against masking, soft elimination, and the generic control all include zero (Supplement N).

An alternative construction uses sixteen sampled Qwen3-8B reasoning chains per item, but only 84.4% parse, below the planned 90% validity floor; 37.5% hit the token cap. Its scripted avoid-commit gain over re-prompting is $+0.6$ points $[-1.5, +2.8]$. Different sampling budgets and truncation prevent a matched-compute interpretation. This exploratory result does not validate the main option-letter measurement (Supplement M).

5 Discussion

The main finding is the separation between a protocol's selection information and a decoder's ability to use it. Commitment sparing under conclusive feedback necessarily creates a Monty gap, yet the tested critics do not establish a reliable gain over re-prompting. They may violate commitment sparing, make false eliminations, or follow a policy different from the decoder's fitted family. The experiments do not identify one universal cause of failure. They show why a correct likelihood-ratio rule is insufficient evidence for a useful intervention.

The theoretical result is useful for inspecting a protocol before fitting a decoder. A fixed lockstep suite can justify masking structurally. A conclusive commitment-sparing interface cannot. But there is no validated deployment example here that combines a useful elimination constraint, reliable verdicts, and an estimable selection policy. If an answer-aware verifier is free to state the answer, these experiments give no reason to prefer selective elimination. We do not conclude that usable selection evidence is impossible or rare across all existing systems.

A discriminating next experiment would randomly assign the same questions to the current critic and a constrained protocol that samples uniformly among its judged-wrong alternatives while sparing the commitment. Actual truth access would remain excluded. Compliance, false eliminations, and rank frequencies would need measurement, with policy fitting and calibration kept separate from evaluation and a predeclared comparison against masking and re-prompting. Uniform sampling over a critic's judged-wrong set is not uniform sampling over the genuinely wrong set. That intervention has not been run here.

The artifact supports auditing the narrower claims: it includes independent decoder implementations, finite-case numerical certificates, reproduction gates, fixed design documents, null readouts, and operator deviations (Supplement F). These checks address implementation and reporting errors; they do not establish general proof validity, interval coverage, or deployment value. Small subject counts, fragile probability measurements, conditional bootstrap fits in the older studies, and the failed reasoning-run validity floor remain substantive limitations. The paper therefore offers a protocol analysis and an empirical boundary on the tested decoders, not a general accuracy-improvement method.

AI USE STATEMENT

Generative AI tools, principally Claude (Anthropic) and Codex (OpenAI), assisted research ideation, mathematical claims and proofs, literature discovery, implementation and experiment execution, synthetic simulations, analysis, figures, and drafting and editing manuscript sections. Feedback from other LLMs, including Grok and Gemini, also informed editorial review. Claude Code operated the GPU dispatches in Supplements L and M; their operator notes record deviations. Codex designed and implemented Supplement N's experiment, which ran on an author-operated RTX 5090, and reproduced its returned analysis. The models studied experimentally are identified in the corresponding methods.

The research question and Monty Hall framing originated with the author. Checks of AI-assisted work include finite-case numerical certificates, source-code inspection, literature checks, and reproduction of reported results; these do not constitute formal verification of the general proofs. The author has personally reviewed the AI-assisted claims, proofs, references, code, and results, and takes responsibility for the final text, claims, code, and other artifacts, including AI-assisted content.

ETHICS STATEMENT

The study uses synthetic simulations and existing benchmarks, without recruiting human participants. The experiments do not establish that selective criticism reliably improves decisions in deployment. In particular, the escalation costs in Supplement N are modeling assumptions, not measurements of human review or guarantees for high-stakes use.

REPRODUCIBILITY STATEMENT

The main text and Appendix A give the central derivations; Appendix B specifies the decoders, estimators, and resampling. The technical supplement retains full measurements and proof extensions under labels A–N, with reproduction commands and finite-case certificates in Supplement F. The artifact includes saved beliefs, critic reveals, code behavior matrices, experiment designs, and operator notes. The confidence experiment's inputs, checkpoints, and analysis are in `experiments/confidence/`. The README distinguishes historical runs from checks repeated for this revision.

REFERENCES

[1] S. Selvin. "A problem in probability" (letter). The American Statistician, 29(1):67, 1975. The name appears in his follow-up, ibid. 29(3):134.

[2] J. S. Rosenthal. "Monty Hall, Monty Fall, Monty Crawl." Math Horizons, 16(1):5–7, 2008.

[3] C. Ma and X. Du. "POE: Process of Elimination for Multiple Choice Reasoning." EMNLP, 4487–4496, 2023.

[4] N. Balepur, S. Palta, and R. Rudinger. "It's Not Easy Being Wrong: LLMs Struggle with Process of Elimination Reasoning." Findings of ACL, 10143–10166, 2024.

[5] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. "Training Verifiers to Solve Math Word Problems." arXiv:2110.14168, 2021.

[6] A. Solar-Lezama, L. Tancau, R. Bodík, S. A. Seshia, and V. A. Saraswat. "Combinatorial sketching for finite programs." ASPLOS XII, 404–415, 2006.

[7] D. F. Heitjan and D. B. Rubin. "Ignorability and coarse data." The Annals of Statistics, 19(4):2244–2253, 1991.

[8] P. D. Grünwald and J. Y. Halpern. "Updating probabilities." JAIR, 19:243–278, 2003.

[9] G. Shafer. "Conditional probability" (with discussion). International Statistical Review, 53(3):261–277, 1985.

[10] P. R. Milgrom. "Good news and bad news: representation theorems and applications." The Bell Journal of Economics, 12(2):380–391, 1981.

[11] K. Park, J. Wang, T. Berg-Kirkpatrick, N. Polikarpova, and L. D'Antoni. "Grammar-Aligned Decoding." NeurIPS, 2024.

[12] Y. Leviathan, M. Kalman, and Y. Matias. "Fast inference from transformers via speculative decoding." ICML, 2023.

[13] E. Botta, Y. Li, A. Mehta, J. T. Ash, C. Zhang, and A. Risteski. "On the query complexity of verifier-assisted language generation." ICML, 2025.

[14] I. Csiszár. "I-divergence geometry of probability distributions and minimization problems." The Annals of Probability, 3(1):146–158, 1975. Thm. 3.1 / Cor. 3.1 give the equality for linear families on a finite alphabet.

[15] H. P. Grice. "Logic and conversation." In P. Cole and J. L. Morgan, eds., Syntax and Semantics 3: Speech Acts, 41–58. Academic Press, 1975.

[16] D. Blackwell. "Equivalent comparisons of experiments." The Annals of Mathematical Statistics, 24(2):265–272, 1953.

[17] S. Kadavath et al. "Language Models (Mostly) Know What They Know." arXiv:2207.05221, 2022.

[18] C. Zheng, H. Zhou, F. Meng, J. Zhou, and M. Huang. "Large Language Models Are Not Robust Multiple Choice Selectors." ICLR, 2024.

[19] P. Pezeshkpour and E. Hruschka. "Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions." Findings of NAACL, 2006–2017, 2024.

[20] N. Balepur, A. Ravichander, and R. Rudinger. "Artifacts or Abduction: How Do LLMs Answer Multiple-Choice Questions Without the Question?" ACL, 2024.

[21] X. Wang, B. Ma, C. Hu, L. Weber-Genzel, P. Röttger, F. Kreuter, D. Hovy, and B. Plank. "'My Answer is C': First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models." Findings of ACL, 2024.

[22] E. A. Cator. "On the testability of the CAR assumption." The Annals of Statistics, 32(5):1957–1980, 2004.

[23] R. D. Gill and P. D. Grünwald. "An algorithmic and a geometric characterization of coarsening at random." The Annals of Statistics, 36(5):2409–2422, 2008.

[24] M. Jaeger. "Ignorability in statistical and probabilistic inference." JAIR, 24:889–917, 2005.

[25] A. V. Gnedin. "The Mondee Gills Game." The Mathematical Intelligencer, 34(1):34–41, 2012.

[26] M. Kleber. "The best card trick." The Mathematical Intelligencer, 24(1):9–11, 2002.

[27] G. J. Simmons. "The prisoners' problem and the subliminal channel." CRYPTO '83, 51–67, 1984.

[28] J. B. Miller and A. Sanjurjo. "A bridge from Monty Hall to the hot hand: the principle of restricted choice." Journal of Economic Perspectives, 33(3):144–162, 2019.

[29] J. P. Morgan, N. R. Chaganty, R. C. Dahiya, and M. J. Doviak. "Let's make a deal: the player's dilemma." The American Statistician, 45(4):284–287, 1991.

[30] Y. Pu, K. Ellis, M. Kryven, J. Tenenbaum, and A. Solar-Lezama. "Program Synthesis with Pragmatic Communication." NeurIPS, 13249–13259, 2020.
[31] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. "Measuring Massive Multitask Language Understanding." ICLR, 2021.
[32] A. Yang et al. (Qwen Team). "Qwen2.5 Technical Report." arXiv:2412.15115, 2024.
[33] A. Grattafiori et al. (Llama Team). "The Llama 3 Herd of Models." arXiv:2407.21783, 2024.
[34] A. Q. Jiang et al. "Mistral 7B." arXiv:2310.06825, 2023.
[35] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen. "MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark." NeurIPS Datasets and Benchmarks, 2024.
[36] D. V. Lindley. "On a measure of the information provided by an experiment." The Annals of Mathematical Statistics, 27(4):986–1005, 1956.
[37] H. Vishwakarma, A. Mishler, T. Cook, N. Dalmasso, N. Raman, and S. Ganesh. "Prune 'n Predict: Optimizing LLM Decision-making with Conformal Prediction." ICML, 61601–61634, 2025. First circulated as "Monty Hall and Optimized Conformal Prediction to Improve Decision-Making with LLMs," arXiv:2501.00555v1.
[38] J. Huang, X. Chen, S. Mishra, H. S. Zheng, A. W. Yu, X. Song, and D. Zhou. "Large Language Models Cannot Self-Correct Reasoning Yet." ICLR, 2024.
[39] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. Bowman, E. Durmus, Z. Hatfield-Dodds, S. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez. "Towards Understanding Sycophancy in Language Models." ICLR, 2024.
[40] M. C. Frank and N. D. Goodman. "Predicting Pragmatic Reasoning in Language Games." Science, 336(6084):998, 2012.
[41] Z. Zhu, B. Liu, Q. Ai, and Y. Liu. "Option-ID Based Elimination For Multiple Choice Questions." arXiv:2501.15175, 2025.
[42] Y. Li, D. Choi, J. Chung, N. Kushman, J. Schrittwieser, R. Leblond, T. Eccles, J. Keeling, F. Gimeno, A. Dal Lago, T. Hubert, P. Choy, C. de Masson d'Autume, I. Babuschkin, X. Chen, P.-S. Huang, J. Welbl, S. Gowal, A. Cherepanov, J. Molloy, D. J. Mankowitz, E. Sutherland Robson, P. Kohli, N. de Freitas, K. Kavukcuoglu, and O. Vinyals. "Competition-level code generation with AlphaCode." Science, 378(6624):1092–1097, 2022.
[43] B. Chen, F. Zhang, A. Nguyen, D. Zan, Z. Lin, J.-G. Lou, and W. Chen. "CodeT: Code Generation with Generated Tests." ICLR, 2023.
[44] F. Shi, D. Fried, M. Ghazvininejad, L. Zettlemoyer, and S. I. Wang. "Natural Language to Code Translation with Execution." EMNLP, 3533–3546, 2022.
[45] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. "Self-Consistency Improves Chain of Thought Reasoning in Language Models." ICLR, 2023.
[46] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. "Let's Verify Step by Step." ICLR, 2024.
[47] E. Kamenica and M. Gentzkow. "Bayesian Persuasion." American Economic Review, 101(6):2590–2615, 2011.
[48] T. M. Cover and J. A. Thomas. Elements of Information Theory, 2nd ed. Wiley, 2006. Theorem 2.7.4 (convexity of mutual information in the channel).
[49] J. Rosenhouse. *The Monty Hall Problem: The Remarkable Story of Math's Most Contentious Brainteaser*. Oxford University Press, 2009. Chapter 4: Progressive Monty.

A FURTHER DERIVATIONS

Scope and notation. These derivations use the full-support subjective joint $P_p(Y, R) = p(Y)L(R \mid Y)$ and conclusive reports. They do not assume that model output probabilities are calibrated to benchmark answers. The separate technical supplement retains the complete historical proofs, certificates, measurements, and deviations under labels A–N. Its labels are prefixed by "Supplement" throughout the main paper.

Deterministic block protocols. Let $F(y)$ be the ordered length-k reveal sequence when the truth is y. For any observed r, set $A(r) = F^{-1}(r)$. Conclusiveness gives $A(r) \subseteq S(r)$. The deterministic likelihood is $\mathbf{1}\{F(y) = r\}$, so the posterior is p restricted to $A(r)$; masking is p restricted to $S(r)$. Because every candidate has positive prior mass, these distributions agree exactly when $A(r) = S(r)$. The nonempty fibers of F partition $\mathcal{Y}$, and every such fiber must then have size $N - k$. Conversely, choose any partition into blocks of that size and assign each block a fixed ordering of its complement. Its fibers equal its survivor sets, so it is ignorable. This proves the deterministic part of Theorem 1, including the divisibility condition.

For completeness, with $m = N/(N-k)$, the number of such protocols is $N!(k!)^m/[((N-k)!)^m m!]$. The factors count partitions into unlabelled blocks and an independent complement order for each block. Ignorable stochastic kernels have the form $L(r \mid y) = \kappa_r \mathbf{1}\{y \in S(r)\}$. Their dimension is $N!/(N-k)! - N$: there is one coordinate per ordered sequence and N independent row-sum constraints. To see independence, if $\sum_{y \in S} a_y = 0$ for every size-$(N-k)$ survivor set, exchanging one member of S shows all a_y are equal, then any set forces that common value to zero. The uniform-among-wrong kernel has every coordinate positive, so nonnegativity does not reduce the affine dimension. Supplement J retains enumeration results and the full extension to adaptive sequences.

Uniform commitment-sparing likelihood. For a legal sequence r sparing c, its likelihood if $Y = c$ is

$$L(r \mid c) = \prod_{t=0}^{k-1} \frac{1}{N-1-t}.$$

For a surviving $y \neq c$, its likelihood if $Y = y$ is $\prod_{t=0}^{k-1}(N-2-t)^{-1}$, since both y and c must be spared. Their ratio telescopes to $(N-1)/(N-1-k)$, proving Proposition 2. The same multiplier applies to every non-commitment survivor; relative ranks within that group do not change. With temperature scaling, the raw ratio is raised to $1/T$, yielding Corollary 2. These calculations require the stated sampling kernel, not merely a commitment-sparing constraint.

Eager decisions. Index candidates in decreasing prior order, with fixed tie-breaking, and suppose the truth has rank j. If $j \leq k$, the eager sequence skips j within ranks $1, \ldots, k+1$ and identifies it. The same j is also the highest-ranked survivor, so masking agrees. If $j > k$, the revealed sequence is $1, \ldots, k$, whose likelihood is constant across survivors. Both decoders then choose rank $k+1$. They agree everywhere and are correct exactly when $j \leq k+1$, proving Theorem 4. There can still be a positive posterior gap: a skip identifies one survivor while masking gives positive probability to other survivors.

Subjective decision bound. For each report, let a_q and a_m maximize q and m. Since $m(a_q) \leq m(a_m)$,

$$0 \leq q(a_q) - q(a_m) \leq |q(a_q) - m(a_q)| + |q(a_m) - m(a_m)| \leq 2\|q - m\|_{\mathrm{TV}}.$$

Pinsker's inequality and Jensen's inequality give

$$0 \leq \delta_p \leq \mathbb{E}_R \sqrt{2 D_{\mathrm{KL}}(q_R \| m_R)} \leq \sqrt{2G}.$$

This proves Proposition 4 in probability units under P_p. It gives no converse, as the eager example shows, and no guarantee on benchmark accuracy when the model belief differs from the answer-generating distribution.

Why a discrepancy can disappear. For a deterministic adaptive protocol, let A_t be the truths compatible with the observed prefix, S_t the candidates not yet revealed, and $q_t = p(\cdot \mid A_t)$. Define the one-step masking update from the current posterior by

$$\widetilde{m}_t(y) = \frac{q_{t-1}(y)\mathbf{1}\{y \in S_t\}}{q_{t-1}(S_t)}, \qquad e_t = -\log q_{t-1}(S_t), \qquad g_t = D_{\mathrm{KL}}(q_t \| \widetilde{m}_t).$$

Here $A_0 = \mathcal{Y}$, so $q_0 = p$. The prefix nesting and conclusiveness imply $A_t \subseteq A_{t-1} \cap S_t$. Both restrictions are normalized, and direct substitution yields

$$e_t + g_t = \log \frac{p(A_{t-1})}{p(A_t)}, \qquad \sum_{t=1}^{k}(e_t + g_t) = -\log p(S_k) + \log \frac{p(S_k)}{p(A_k)}.$$

The final logarithm is $D_{\mathrm{KL}}(q_k \| m_k)$. If the final fiber equals the survivor set, it vanishes even though earlier g_t can be positive. Selection then told the decoder early what later eliminations told masking. This is the sense in which cancellation is anticipation, not the loss of already acquired information. Supplement J, Proposition S4, gives the complete statement and numerical certificates.

B DECODER AND POLICY ESTIMATORS

The following specification describes the saved implementations. Historical code uses the original proposition numbers; the concordance at the front of the technical supplement maps them to this paper. The conclusive likelihood update is Bayesian when its assumed policy is correct. Fallbacks, empirical tables, and the positive-error extension are separate implementation conventions.

Algorithm 1: sequential decoding with masking fallback

1. Inputs: full-support p, fixed commitment c, observed distinct valid reveals, assumed policy π, hedge $\epsilon \in [0, 1]$, and uniform-error mixture weight $\eta \in [0, 1]$. Initialize weights $w_y = p(y)$ and revealed set $H = \varnothing$.

2. For each valid reveal r, let $U = \mathcal{Y} \setminus H$. Compute the conclusive policy likelihood $\ell_\pi(y) = \pi(r \mid y, c, p, H)$ and the uniform-among-wrong likelihood $\ell_0(y) = \mathbf{1}\{r \neq y\}/(|U| - 1)$ on compatible histories. Set

$$\ell_\epsilon(y) = (1 - \epsilon)\ell_\pi(y) + \epsilon\ell_0(y), \qquad \ell_{\epsilon,\eta}(y) = (1 - \eta)\ell_\epsilon(y) + \eta/|U|.$$

   Update $w_y \leftarrow w_y \ell_{\epsilon,\eta}(y)$ and $H \leftarrow H \cup \{r\}$. In the hard branch ($\eta = 0$), set $w_r = 0$. Missing or invalid verdicts leave w, H unchanged.

3. At the evaluated budget, compute $m_y = p(y)\mathbf{1}\{y \notin H\}/\sum_{z \notin H} p(z)$. If $\sum_y w_y = 0$, return $q = m$; otherwise return $q = w/\sum_y w_y$. Return the argmax using the implementation's fixed tie-breaking rule.

The contradiction test is zero total weight at the evaluated budget, not a disagreement with the argmax, a small likelihood, or a per-round reset. Once all weights are zero they remain zero, and the final output is masking on *all* observed valid reveals. The default is $\epsilon = \eta = 0$. The hedge and error components are composed in that order in the formula; the saved hedge sweeps and critic error analysis test the components separately, not a new joint sweep.

The conclusive branch is `mhd_sim.py:run_episode`; critic replay is `gpu_experiments/critic/analyze.py:replay`. The latter uses denominator floors of 10^{-300}, skips invalid verdicts, and supports a soft branch without hard zeroing. In the soft assumed-policy branch, the code evaluates its likelihood routine on the observed history even for hypotheses contradicted earlier. This is an implemented robustness extension, not a fully normalized generative model of every off-support history. The conclusive theorems do not apply to that branch. Also, η is a mixture weight: uniform sampling over U hits the truth with probability $\eta/|U|$ if it is still in U. The measured critic false-elimination rate is plugged into η heuristically, not inverted to recover a generative mixture rate.

Key-trained tables. For item i, let c_i be the commitment and let ranks use decreasing p_i with stable index tie-breaking. Under a hypothesized truth $y \neq c_i$, label each possible reveal by one of N roles: TRUTH, STATED, or OTHER$_j$, where $j = 1, \ldots, N-2$ orders the candidates other than y, c_i by their student rank. Under $y = c_i$, the roles are TRUTH=STATED and SELF$_j$, $j = 1, \ldots, N-1$, again ordered by rank. Let $v_i(y) \in \{\text{split}, \text{same}\}$ identify the table and $a_i(y, r)$ the role. Each vocabulary is a bijection between options and roles.

For each reveal step separately and each training fold, count the observed role using the training answer key. If $C_{v,a}$ is its count among valid verdicts, fit

$$\widehat{\theta}_{v,a} = \frac{C_{v,a} + 1}{\sum_b C_{v,b} + N}.$$

The evaluation item's likelihood is $\widehat{\theta}_{v_i(y),a_i(y,r)}$. At later steps the table is renormalized over unrevealed options for each hypothesis. Two folds are assigned by a seed-7 random permutation of item indices modulo two; each item is decoded only with the opposite fold's fitted tables. Pooled tables are descriptive only. The primary estimated decoder uses hard elimination; the reported soft variant omits it.

Rank flattening. Keep the split table's TRUTH and STATED cells fixed and replace all OTHER cells by their arithmetic mean. Keep the same table's TRUTH=STATED cell fixed and replace all SELF cells by their mean. This preserves each table's sum and removes distinctions between these ranks. It retains the truth and commitment rates, including their training-key information. Thus it is a rank-preference control, not a key-free likelihood or proof that all critic competence has been removed.

Log-only constrained EM. The primary log-only estimator uses the same role map but no training answer keys. At the first step set $b_i(y) = p_i(y)$; at later steps mask previous reveals in p_i and renormalize. Start both tables uniformly over their non-truth cells, with the truth cells clamped to zero. For each valid training item, iterate

$$z_i(y) = \frac{b_i(y)\theta_{v_i(y),a_i(y,r_i)}}{\sum_u b_i(u)\theta_{v_i(u),a_i(u,r_i)}}, \qquad C_{v,a} = \sum_{i,y} z_i(y)\mathbf{1}\{v_i(y) = v,\ a_i(y,r_i) = a\}.$$

If the first denominator is zero, use $z_i = b_i$. Add one to each allowed count, set the truth cells to zero, and normalize each table. Stop when the largest cell change is below 10^{-10} or after 5,000 iterations. Flatten the fitted non-truth rank cells after fitting, then decode the opposite fold. The code uses seed-7 folds. This is smoothed constrained EM for the first-step model; later steps are per-step plug-in fits using masked priors, not maximum likelihood for the entire sequential process. The assumptions are calibration of b_i, the two-table policy family, and conclusive reports. Real false eliminations violate the last assumption.

The code also reports unconstrained EM, constrained moment matching, and a signature-only estimator. These are secondary arms. The unconstrained fit estimates truth cells; moment matching minimizes squared reveal-indicator error on the same constrained simplices; the signature estimate is $\widehat{s} = \text{mean}\,\mathbf{1}\{r = c\}/\,\text{mean}(1 - p(c))$ on valid training items. Source: `code/tier2/logonly.py`; the primary arm is `emc_flat`.

Donor-rank resampling control. The quantity previously called a shuffle or permutation p-value uses 200 *resampling* replicates (seed 11). At each step, rank the non-truth candidates of each item. Keep missing verdicts and false eliminations unchanged. For each remaining item, sample with replacement a donor rank from the pool of valid non-truth reveals at that step, and strike the recipient's non-truth candidate at that rank. If this would repeat an earlier reveal, retain the original reveal. Refit the tables, re-decode, and compare the absolute decoder-minus-masking gain with the observed absolute gain. The reported value is $(1 + \#\{\text{replicate gains at least as large}\})/201$, with minimum $1/201$.

This preserves false-elimination indicators and samples from the observed donor-rank distribution; it does not preserve finite-sample rank counts exactly and is not an exact permutation test. Its calibration as a null test requires a suitable exchangeability assumption. We therefore treat it as a diagnostic control and base the main empirical conclusions on paired differences, intervals, and replication across configurations.

Intervals. Item intervals are paired percentile bootstraps of item correctness differences. They resample the saved predictions without refitting the policy estimator; uncertainty in the fitted tables is therefore not fully propagated. Cluster intervals resample subject labels with replacement, concatenate all items in the sampled labels, and recompute the item-weighted mean; percentiles are 2.5 and 97.5. Ten or twelve observed clusters provide limited information about subject-level generalization. The saved scripts generally use 10,000 replicates; per-experiment seeds and exceptions are recorded in their analysis code. The sham-key subset is the first 1,000 items in file order, with prior accuracy 62.4% versus 55.2% on the full pool, so cross-subset differences are not a clean treatment comparison.