

Article

Question Selection and Answer Scoring Shape Revision Risk in Recurrent Language Models

Ruian Lei ¹, Yazhou Ren ¹ and Zenglin Xu ^{2,*}

¹ School of Computer Science and Engineering, University of Electronic Science and Technology of China, Qingshuihe Campus, No. 2006 Xiyuan Avenue, West Hi-Tech Zone, Chengdu 611731, China; ryanl057@std.uestc.edu.cn (R.L.); yazhou.ren@uestc.edu.cn (Y.R.)

² Artificial Intelligence Innovation and Incubation Institute, Fudan University, Zhangjiang Campus, No. 825 Zhangheng Road, Pudong, Shanghai 201203, China

* Correspondence: zenglinxu@fudan.edu.cn

Abstract

An evaluation of early answers must choose which questions to release and how to read their answers. Using one scoring convention for both choices can make two procedures look similar while concealing their sensitivity to either choice. We measured this dependence on stored trajectories from Huginn and Ouro, using cloze scoring on three filtered benchmarks with 200 questions per task and cohort. The historical cohorts share questions; the three cohorts are not independent replications. At 50% release in the headline windows, changing the selection margin with the answer readout fixed changed revision risk by 9.33 to 12.00 percentage points (pp). Changing the readout with the margin fixed changed it by 9.67 to 13.00 pp. Yet changing both together gave only 0.33 to 3.00 pp. Similar risks for complete procedures therefore did not establish interchangeable selection margins. Separate sum-to-mean grids within each length unit also placed the lowest tested risks near matching rules. At the headline windows, final-depth disagreement retained the fixed-readout margin effect, although its magnitude depended on the outcome definition and observation window. Lower revision risk established neither an accuracy gain nor better stopping decisions. The results support a practical evaluation check: hold the answer rule fixed when comparing selectors, then cross the rules to measure dependence on the target.

Keywords: adaptive computation; recurrent reasoning; early exit; answer scoring; surprisal; candidate entropy; revision risk; selective prediction; evaluation protocols

1. Introduction

When a recurrent language model performs more computation, its preferred answer can change. An evaluator may want to return some answers early and measure how often they would later change. Four terms describe this evaluation. The *readout* converts candidate scores into an answer, usually by choosing the highest score. The *margin* is the gap between the two highest scores; it ranks questions by the lead of the current winner. A *release* returns an answer at an earlier depth. A *revision* occurs if that answer changes at any observed later depth, including a change followed by a return. Here, release is evaluated on stored trajectories rather than deployed as a stopping policy.

The same scoring convention often supplies both the margin and the readout. This couples two choices: which questions enter the released set, and what counts as a revision on those questions. Comparing two such complete procedures changes both at once. Similar reported risks therefore do not show that their margins can be used interchangeably



Academic Editor: Christos Bouras

Received: 15 September 2026

Revised: 1 October 2026

Accepted: 1 October 2026

Published: 5 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

under a fixed answer rule. We measure the size of this distinction by crossing selection margins and answer readouts on the same model outputs.

For example, on our new Ouro questions, the token and byte procedures gave revision risks of 13.00% and 13.67% when each used its own margin to release half the questions. Keeping the byte answer rule but selecting with the token margin raised risk to 25.67%. The margins therefore selected sets with substantially different risks under the same answer rule, despite the similar risks of the complete procedures.

The direction of the dependence is plausible before seeing the data. A large margin under convention w means that w 's own winner has a substantial lead. Selecting on that margin should tend to protect answers defined by w more than answers defined by another convention v . This is an expectation, not a guarantee: an anchor margin alone does not constrain subsequent score changes. Nor does it determine the size of the risk difference. In our headline comparisons, changing one rule shifted risk by roughly 10 percentage points (pp), whereas changing both shifted it by only 0.33 to 3.00 pp. The contribution is this measured dependence and its consequence for interpreting evaluations. The algebra that separates the two changes is standard.

Length normalization already affects candidate preferences and accuracy. OLMES compares scoring conventions and makes task-specific recommendations [1]; Oostermeijer analyzes length bias and proposes a Bayesian correction [2]. Surface form competition motivates other corrections, including the pointwise mutual information (PMI) scoring that OLMES recommends for our three tasks [1,3]. Presenting answer options together instead of scoring separate continuations also changes measured ability [4]. Changes in answer selection can move model rankings [5]. The Language Model Evaluation Harness paper describes the difficulty of reproducing these choices [6]; its documentation and a repository issue record the character-versus-byte normalization distinction used here [7,8]. We extend this question from predictions at one depth to revision events across depths and to the questions selected for evaluating them.

Prediction churn distinguishes changes in individual predictions from changes in aggregate performance [9]. Agreement between measurement summaries likewise need not imply agreement on individual cases [10,11]. Kim and Kang show that confidence comparisons can change when the scoring context changes while generated answers and correctness labels remain fixed [12]. Our comparison fixes candidate forward outputs and varies their arithmetic readout. It separates risk changes caused by selecting different questions from changes caused by redefining revision.

Adaptive computation methods learn when to stop or select exits at inference [13–17]. Confident adaptive language modeling (CALM) and risk-controlled early exit evaluate explicit consistency or performance objectives [18,19]. Our release fraction and revision risk form a risk–coverage pair, as in selective classification [20,21]. Adaptive-depth evaluations of looped models also separate trajectory formation from exit selection [22]. We study a further choice within the evaluation: the answer rule defining the revision objective. Crossing selectors and answer rules identifies which choice accounts for the measured risk difference.

We use Huginn and Ouro trajectories from three filtered multiple-choice benchmarks, with 200 questions per task in each of three cohorts [23,24]. Two historical cohorts share questions across models; a later Ouro cohort contains questions not previously collected in this study. All principal comparisons use cloze scoring and 50% release. We report crossed risks, task-level results, and sensitivity to the revision outcome and observation window. Separate exponent grids test sum-to-mean families within a length unit. A supporting two-point check tests whether normalization can change a winner at a fixed depth. Together

these analyses give evaluators a way to expose dependence that a comparison of complete procedures can conceal.

2. Materials and Methods

2.1. Trajectories, Cohorts and Scoring Rules

We used the released Huginn checkpoint `tomg-group-umd/huginn-0125` and Ouro checkpoint `ByteDance/Ouro-2.6B` without further training. The tasks were ARC-Challenge [25], CommonsenseQA [26] and OpenBookQA [27]. Each historical task cohort contained 256 source questions and 200 disjoint evaluation questions. The historical evaluation questions were shared across the two models. Huginn outputs had already been inspected, so analysis of its historically named “confirmation” cohort is retrospective. The original Ouro replication collected new outputs on previously studied questions after its protocol was fixed. A later Ouro cohort added 200 questions per task under a separate local prospective protocol. These filtered benchmark subsets support comparisons within the sampled questions and models, with limited coverage of full benchmarks and model architectures. *New questions* means new to this study’s recorded collection, not known to be absent from model pretraining. OpenBookQA uses the test split historically and validation in the new cohort.

Huginn’s 16/128 window was inherited from the exploratory analysis of its stored 128-step trajectories. It compares an early observation with a longer observed future; it was not selected by a prospective optimization criterion for this paper. Ouro’s horizon is its default four passes, with pass 2 as an intermediate anchor, fixed before the cloze replication was collected. A Huginn step and an Ouro pass are model-specific recurrence units, not equal units of cost. The revision analysis tests additional anchors and horizons on the same scores rather than treating either headline window as generally optimal. Table 1 summarizes the cohorts, observation windows and evidence status.

Table 1. Evidence units in the main measurement. Each row contains 600 questions, 200 per task. Historical questions overlap between models, and multiple analyses reuse each row’s outputs. The newer OpenBookQA cohort uses validation; the historical cohort uses the test split.

Cohort	Anchor/Horizon (Huginn: Steps; Ouro: Passes)	Evidence Status
Huginn historical	16/128	Exploratory reanalysis of existing outputs
Ouro historical	2/4	Original prespecified replication with new outputs on historical questions
Ouro new questions	2/4	Local prospective factor protocol; later descriptive analyses reuse its outputs

For question i , candidate j and recurrent depth u , let $L_{ij}(u)$ be its teacher-forced continuation log likelihood. Cloze scoring evaluates the complete candidate text after the question prompt. Within each model, all scoring views use the same candidate forward passes, inputs and initialization. They differ only in the arithmetic readout. Huginn used padded candidate batches with a shared initialization draw; Ouro evaluated separate candidate sequences. Invalid candidate slots are masked, and answer ties use the smallest original candidate index. Collection details and numerical acceptance checks for both historical and new-question cohorts are given in Supplementary Section S1. Supplementary Data S1 provides their protocols, numerical metadata and stored scores. Table 2 summarizes the scoring rules and their roles.

Table 2. Scoring rules and their roles. All length divisors are positive. The cloze rules count the scored continuation, including its leading space. H1 is the historical Language Model Evaluation Harness reference described below.

Rule	Score	Role Here
Sum	L_{ij}	Second readout in the paired label comparison
Byte mean	L_{ij}/b_{ij}	Principal readout throughout
Token mean	L_{ij}/k_{ij}	Principal alternative margin
Cloze character mean	L_{ij}/c_{ij}	Convention pairs and invariance checks
H1 character mean	$L_{ij}/c_{ij}^{\text{H1}}$	ARC/OpenBookQA harness reference

Cloze character length c_{ij} counts Unicode code points in the scored continuation. H1 denotes the historical reference using public Language Model Evaluation Harness task definitions [6]. It scores answer texts by character-normalized likelihood on ARC and OpenBookQA, but answer letters by unnormalized likelihood on CommonsenseQA. For its character-normalized scores (acc_norm), c_{ij}^{H1} counts the original choice without the target delimiter. These are different text boundaries. The issue and metrics documentation identify the corresponding implementation distinction [7,8].

Write $s_{ij}^v(u)$ for a score and $y_i^v(u) = \arg \max_j s_{ij}^v(u)$ for its returned answer. The negative byte/token mean scores are average surprisal in different units. Candidate entropy is a separate quantity, obtained from $p_{ij}^v = \exp(s_{ij}^v/\tau) / \sum_{j'} \exp(s_{ij'}^v/\tau)$. The temperature τ sets its scale. This candidate-restricted softmax has no established calibration as a correctness probability.

2.2. Revision Labels, Margins and Readouts

For anchor t and observed horizon T , define

$$R_i^v = \mathbf{1}\{\exists u \in \{t+1, \dots, T\} : y_i^v(u) \neq y_i^v(t)\}. \quad (1)$$

A revision followed by return to the anchor answer counts as a revision. The final-depth alternative is $R_{i,\text{end}}^v = \mathbf{1}\{y_i^v(T) \neq y_i^v(t)\}$. Transient revisions have $R_i^v = 1$ and $R_{i,\text{end}}^v = 0$. Correctness against gold g_i is a separate outcome. Later observations are used to evaluate outcomes; margins use only the observed prefix. For byte/sum revision labels, we report the complete paired 2×2 table. If its counts are $n_{00}, n_{01}, n_{10}, n_{11}$, positive specific agreement is $2n_{11}/(2n_{11} + n_{01} + n_{10})$ and negative specific agreement is $2n_{00}/(2n_{00} + n_{01} + n_{10})$. We also report Cohen's κ , whose dependence on marginal prevalence limits its use as a standalone measure of measurement validity [28,29].

Let $A_i^w \in [0, 1]$ denote release using the top-two margin under margin rule w at depth t . The principal fixed-quota comparison releases the highest-margin fraction $q = 0.5$ in each task: 100 of 200 evaluation questions. At boundary ties, fractional membership is the expected inclusion under uniform random selection to meet the quota exactly: weights are 1 above the boundary, 0 below it, and equal fractional weights at the boundary whose sum fills the remaining quota. Thus A_i^w is an inclusion weight rather than always a binary indicator. Define

$$r(v, w) = \frac{\sum_i A_i^w R_i^v}{\sum_i A_i^w}, \quad D(v \leftarrow w) = r(v, w) - r(v, v). \quad (2)$$

Here v is the readout, which fixes the returned answer and the revision event, and w is the margin, which fixes what is released. Both estimands are valid for their respective questions. A fixed readout need not supply the best margin.

Call $r(b, b)$ and $r(k, k)$ the matched settings, each taking its margin and its readout from one convention, and $r(b, k)$ the crossed setting. Then

$$\underbrace{r(b, b) - r(k, k)}_{\text{complete procedure}} = \underbrace{(r(b, k) - r(k, k))}_{\text{readout leg}} - \underbrace{(r(b, k) - r(b, b))}_{\text{margin leg}}, \quad (3)$$

so comparing complete procedures returns the difference of two legs that each measure one substitution against the crossed setting. Read as a walk from $r(k, k)$ to $r(b, b)$ through that setting, the first step adds the readout leg and the second subtracts the margin leg.

Between the sum rule and the length mean, Equation (6) below supplies a continuous family. We therefore computed $r(\alpha, \beta)$, the risk under the readout at exponent α among the questions released by the margin at exponent β , on the grid $\alpha, \beta \in \{0, 0.1, \dots, 1\}$, separately for the token and byte units. The token grid crosses sum and token mean, and the byte grid crosses sum and byte mean; the byte/token crossed settings belong to neither grid. The quota, tie handling and bootstrap draw are those of the legs, so all 121 cells of one cohort share one resample. Three criteria were fixed before any interior cell was computed, after the corner results were known. The first required the mean risk of cells whose exponents differ by at least 0.5 to exceed the diagonal mean, with an interval excluding zero, in every cohort and unit. The second required the mean risk at each exponent distance to increase with distance, with Spearman correlation at least 0.8 and a positive difference between the farthest and the middle distance. The third required the margin exponent that minimises risk at a fixed readout to lie within 0.2 of the readout exponent in at least 9 of the 11 rows. These summaries used the smallest exponent attaining the row minimum. The registered analysis also recorded whether this chosen minimiser was nearly constant. A post-result check tested for a common empirical minimiser across all readouts, retaining every exponent whose unrounded risk was within 10^{-9} percentage points of its row minimum and intersecting these sets over rows. This tolerance left the original smallest-exponent choices unchanged.

Whole-cohort hybrid accuracy is

$$a^v(A) = \frac{1}{n} \sum_i [A_i \mathbf{1}\{y_i^v(t) = g_i\} + (1 - A_i) \mathbf{1}\{y_i^v(T) = g_i\}]. \quad (4)$$

Revision risk conditions on releases; hybrid accuracy uses all questions. Neither quantity measures latency or floating-point operations (FLOPs).

2.3. Question-Wide and Candidate-Specific Rescaling

We separate two changes in the selection margin while keeping the byte answer rule fixed. The question-wide factor Q scales all candidates in one question equally; the candidate-specific factor C changes their relative scales. The four settings apply neither factor, Q alone, C alone, or both (QC).

Let $r_{ij} = b_{ij}/k_{ij}$, $\bar{r}_i$ its geometric mean within question, and G the geometric mean of $\bar{r}_i$ in the fixed source cohort, separately by task. Put $q_i = \bar{r}_i/G$ and $d_{ij} = r_{ij}/\bar{r}_i$. Then

$$s_{ij}^b = \frac{s_{ij}^k}{Gq_id_{ij}}. \quad (5)$$

The four margins are base s^k/G , $Q s^k/(Gq_i)$, $C s^k/(Gd_{ij})$, and QC s^b , while the readout stays the byte mean. The Q and C contrasts each subtract base, giving simple effects with the other factor absent. Each contrast therefore refers to a fixed setting of the other factor. A positive divisor shared by all candidates in a question preserves their ranking

but can change the ranking of margins across questions. Candidate-specific divisors can change both rankings. The resulting revision-risk contrasts must be measured empirically. The rescaling alone cannot identify a mechanism of recurrent state evolution.

2.4. Testing Normalization Invariance at a Fixed Depth

We now ask whether changing normalization can change the winning answer at one observed depth. Equal divisors within a question guarantee answer agreement across sum and length mean. Agreement can also be checked directly from the winners at both normalization endpoints of a continuous family. These endpoints are the two scoring rules commonly reported by an evaluation harness. For a length unit with $\text{len}_j > 0$ and negative log likelihoods $\ell_j > 0$, write

$$s_j(\alpha) = -\ell_j \text{len}_j^{-\alpha}, \quad \alpha \in [0, 1], \quad (6)$$

so that $\alpha = 0$ is the sum rule and $\alpha = 1$ the length mean.

Proposition 1 (Two-point test). *Suppose a question has a strict maximizer of $s_j(\alpha)$ at $\alpha = 0$ and at $\alpha = 1$. Then one candidate is the strict maximizer for every $\alpha \in [0, 1]$ if and only if the two endpoint maximizers coincide.*

Proof. Because the scores are negative, maximizing them is equivalent to minimizing $\log \ell_j - \alpha \log \text{len}_j$. Differences between these expressions for any two candidates are affine in α . A strict inequality that holds at both endpoints therefore holds throughout the interval. Applying this to each competitor proves sufficiency; necessity follows by evaluating the interval's common strict maximizer at its endpoints. $\square$

Corollary 1 (Several length units). *Let $s_j(\alpha) = -\ell_j \prod_{u=1}^d \text{len}_{ju}^{-\alpha_u}$ for d length units at once. Differences between the candidate expressions $\log \ell_j - \sum_u \alpha_u \log \text{len}_{ju}$ are affine in α , so a common strict maximizer at the 2^d vertices of a hyperrectangle $\prod_u [\alpha_{u0}, \alpha_{u1}]$ is the strict maximizer throughout it.*

The test uses each question's strict endpoint winners. A harness reporting both `acc` and `acc_norm` has already taken the argmaxes at $\alpha = 0$ and $\alpha = 1$. Retaining their identities and checking strictness allows certification without another forward pass. Proposition 1 gives an exact test within this family; Corollary 1 extends it to d units using 2^d vertices. An archive containing correctness labels alone identifies the winner when both labels are correct, but leaves its identity unknown when both are incorrect. The test applies a standard property of affine functions. The maximizing candidate is piecewise constant in the parameter and changes only at pairwise intersections; agreement of the strict endpoint winners therefore certifies the interval [30]. Here the endpoints coincide with scoring rules already used in evaluation. Section 3.6 applies the test to stored winners from these cohorts and public archives.

The certificate is specific to a length unit. Token and character means are different vertices of the rectangle in Corollary 1, so a question certified for the token family can still change its winner between them. Section 3.6 reports each unit separately.

The test uses scores as well as lengths. Oostermeijer gives a rule of thumb that near-equal candidate lengths limit length bias [2]. Exactly equal lengths guarantee winner agreement; near equality remains a heuristic. Proposition 1 gives a necessary and sufficient condition once the scores are in hand. Section 3.6 compares its coverage with the exact equal-length condition and reports near-equal lengths separately as a descriptive screen.

The family is multiplicative. Additive corrections such as pointwise mutual information lie outside it [3]. The test certifies the answer at one observed step. It gives no guarantee against revision at a later depth or against a change in which questions the margin releases.

2.5. *The Models' Own Exit Signals as Margins*

We also used each model's own exit signal at the anchor to select questions. These signals differ from the candidate scores used in the matched setting. The byte mean remained the readout, keeping the revision labels fixed. Huginn's stored trajectories carry its shipped exit criteria at every step, computed on the prompt at its last position with the official evaluators and initialisation. These are the divergence between successive next-token distributions, its min- p variant, the change in entropy, the relative change in the latent state, its cosine distance, and the number of steps over which the next-token argmax has been unchanged. Ouro's learned exit gate was not stored. The frozen model was therefore re-run on each question's prompt alone for its four steps, under the frozen dtype, attention kernel, determinism settings and pinned libraries. The gate's instantaneous and cumulative exit probabilities were recorded at every prompt position, together with the relative hidden-state displacement at the last one. Before any gate was kept, one stored candidate row per task and cohort was replayed bit for bit. The gate at the prompt positions of that candidate sequence was required to equal the gate of the prompt alone. Each signal releases the 50% of questions it marks as most converged, with the same quota, tie handling and bootstrap draw as the other legs. The divergence criterion for Huginn and the cumulative gate for Ouro were registered as primary before computation. We registered two criteria. The first required the releases selected by these signals to carry more byte-mean revision risk than the matched byte setting in at least two cohorts. The second required their risk to be at least as high as in the crossed setting using the token margin. Prompt-only signals decide release before any candidate is scored, which is the order an adaptive-depth pipeline follows.

2.6. *Uncertainty, Provenance and Analysis Status*

The original Ouro primary estimand was the equal mean of three tasks and two byte/token return directions at 50% coverage. It is distinct from the later three-task margin leg. Original intervals use 2000 paired whole-question bootstrap replicates within task. The same draw is shared across views and outcomes, quota rankings are recomputed, and task/direction means are formed before percentile bounds. These intervals assume exchangeable questions within each filtered task sample; they are not population guarantees for arbitrary benchmarks.

Table 3 distinguishes criteria fixed before new collection, criteria fixed on already available data before a particular calculation, and post hoc descriptions. Local hashes record protocol files and collection markers; they are neither external preregistration nor independent proof of chronology. The new-question protocol preceded collection, but its implementation followed collection. Supplementary Section S1 records an amendment to one control. Related historical results were already known.

Table 3. Analysis status and original decision rules. Before computation refers only to the named analysis, not to first access to the data. P: criterion fixed before new output collection; A: fixed before this calculation on available data; H: post hoc or descriptive, with no original pass/fail rule. A later multiplicity assessment does not change these statuses. CI: confidence interval.

Analysis	Timing	Criterion or Purpose	Outcome
Historical Ouro primary	P	Two-direction mean CI above zero; positive task means in $\geq 2/3$ tasks	Passed
New Ouro C effect	P	CI below zero; negative effects in $\geq 2/3$ tasks	Passed
New Ouro Q effect	P	Expected near zero; no equivalence margin	Equivalence unresolved
Byte/sum labels; separate legs; historical factors	H	Agreement and separate substitutions	Descriptive
Convention pairs	A	Both crossed risks above both matched in $\geq 8/9$ cohort–pair cells	Passed (9/9)
Exponent grid	A	Far-minus-diagonal CI above zero; distance correlation ≥ 0.8 and positive far-minus-middle estimate; minimiser within 0.2 in $\geq 9/11$ rows	All 3 passed
All tied grid minima	H	Common minimiser across readouts	None found
Native exit signals	A	(1) Excess over matched in ≥ 2 cohorts; (2) risk at least token-crossed	(1) Passed; (2) failed
Huginn at Ouro window	A	(1) Both legs resolved, matched below crossed; (2) relabelled-set overlap CI above chance	Both failed
Exit-signal ranking	A	Reversal; both contrast CIs exclude zero	Failed
Release scan; accuracy; first crossing; archive and certificate checks	H	Operating-point sensitivity, utility and coverage	Descriptive
Revision sensitivity; multiplicity	H	Scope fixed after original results; no new confirmatory claim	Descriptive

For the revision, we use 20,000 paired resamples and keep the original 2000-draw results unchanged. The additional analyses cover 13 Huginn windows and all six anchor/horizon pairs among Ouro passes 1–4. For each window, the any-time and final-depth outcomes use the same anchor margins and release weights. Historical models share re-sampled question indices where their questions align. Supplementary Section S16 provides the full windows, task-level estimates and transient-revision counts.

We also introduce two explicit multiplicity families in this revision. Family P groups three quantities designated primary or co-primary in their original protocols: the historical Ouro two-direction mean and the new Ouro C and Q effects. These came from separate protocols; their grouping here is new. Family H contains the six post hoc headline margin/readout legs across three cohorts. For a family of size m , Bonferroni percentile intervals use bootstrap quantiles $0.05/(2m)$ and $1 - 0.05/(2m)$. Their family coverage remains approximate because the underlying bootstrap intervals are approximate. Bonferroni does not require independence, but does not turn reused questions into replications or post hoc choices into prespecified tests. Other intervals remain pointwise. The sensitivity exports contain 500 aggregate and 1500 task-level intervals across outcomes, windows, risks and contrasts; they are descriptions, not separate confirmatory findings.

The scalar replay is a separately written arithmetic implementation using the same stored scores and metadata. It did not import the original analysis implementation. It is not independent data, independent authorship or external validation. It reproduced eleven

main effect/accuracy point estimates and three revision-label count tables. A separate numerical port of the original bootstrap algorithm reproduced the original risks, contrasts and intervals, including Table 4. Artificial intelligence (AI) coding assistance was used in the checks; the consolidated disclosure appears in the Acknowledgments. No new model inference was needed for the revision analyses.

Table 4. Revision risk at 50% release. Matched settings use one convention for both rules; the crossed setting uses the token margin with the byte readout. The separate legs are post hoc (Table 3). Contrasts satisfy Equation (3). Pointwise 95% paired percentile bootstrap intervals use 2000 shared draws and are unadjusted; family-wise intervals for the legs are in Supplementary Table S25. Each cohort has 600 questions with equal task weights.

(a) Revision risk among released questions (%)			
Cohort	Matched, byte	Matched, token	Crossed
Huginn, earlier questions	2.33	2.00	11.67
Ouro, earlier questions	13.33	10.33	23.33
Ouro, new questions	13.67	13.00	25.67
(b) Risk difference in percentage points (pp) with 95% intervals			
Cohort	Complete procedure	Margin leg	Readout leg
Huginn, earlier questions	0.33 [−2.00, 2.67]	9.33 [5.33, 12.33]	9.67 [5.33, 13.00]
Ouro, earlier questions	3.00 [−2.33, 7.00]	10.00 [5.33, 14.67]	13.00 [7.32, 17.67]
Ouro, new questions	0.67 [−4.33, 6.67]	12.00 [6.33, 15.33]	12.67 [6.33, 17.34]

3. Results

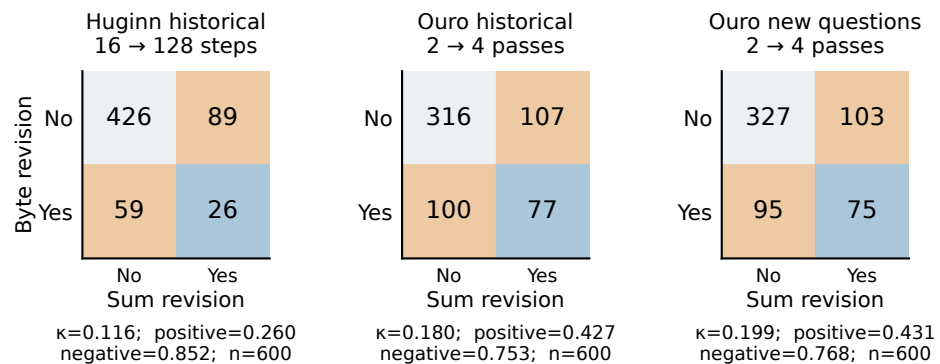
3.1. Paired Revision Labels Reveal Readout Dependence

We first compare revision labels on all questions, without selecting a release set. Byte-mean and sum readouts of the same trajectories identified substantially different sets of revising questions (Figure 1a). For Huginn historical, Ouro historical and Ouro new questions, respectively, the byte-mean revision counts were 85, 177 and 170 of 600; the sum counts were 115, 184 and 178. The corresponding intersections contained only 26, 77 and 75 questions. The two readouts therefore assigned different binary revision labels to 148, 207 and 198 questions. Similar marginal counts can therefore describe different sets of revising questions.

Positive specific agreement ranged from 0.260–0.431, whereas negative specific agreement ranged from 0.753–0.852. Cohen’s κ was 0.116–0.199; its cohort intervals were [0.026, 0.204], [0.092, 0.257] and [0.114, 0.283]. The count tables report the marginal prevalences for each summary. Both readouts were applied deterministically to the same outputs. Their agreement measures consistency between scoring rules; neither provides a validated reference for internal stability. A fixed length divisor preserves temporally constant scores. On evolving trajectories, it can change which candidate crosses a decision boundary.

3.2. Similar Matched Risks Concealed Larger Changes in Each Rule

We next cross byte and token scoring rules to separate question selection from the revision target. The complete procedures compared here take both rules from one convention. At 50% release the two matched settings carried similar risk: 2.33, 13.33 and 13.67% under byte mean, against 2.00, 10.33 and 13.00% under token mean. Using the token margin with the byte readout raised risk to 11.67, 23.33 and 25.67% (Table 4; Figure 1b).

a Revision labels: byte mean vs sum**b Crossed risks: byte mean vs token mean**

Point labels give readout / selection margin; 50% release

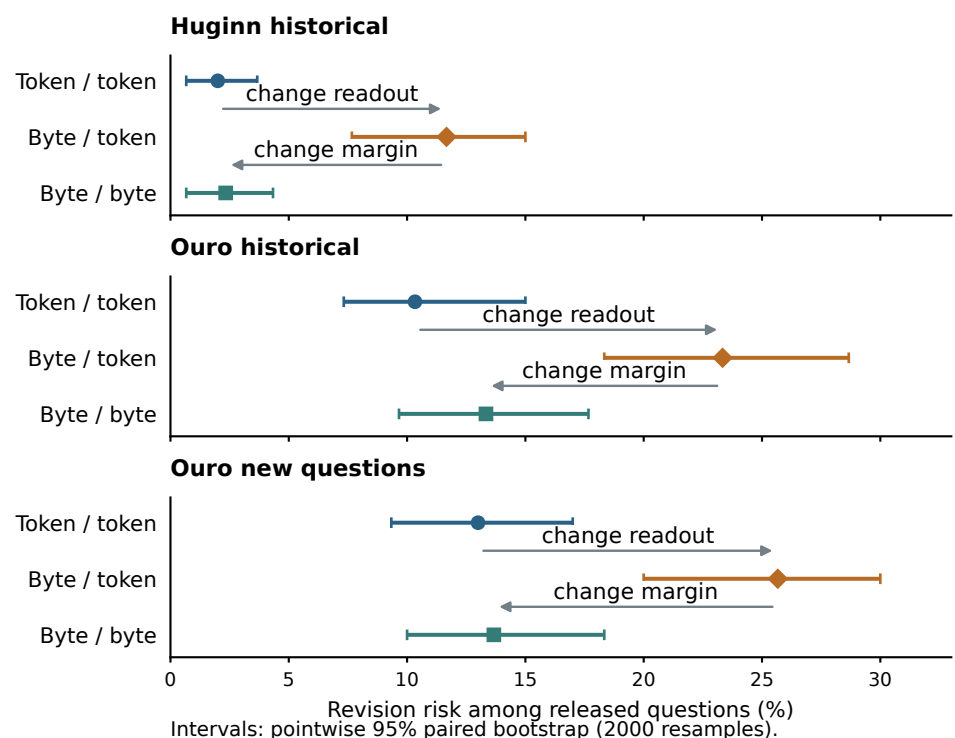


Figure 1. Paired revision labels and separate changes in margin and readout. **(a)** Complete revision-label counts for byte mean versus sum; rows indicate byte revision and columns sum revision. Each cohort has 600 questions. The off-diagonal cells identify different labels; positive and negative specific agreement accompany κ . **(b)** The three risks in Table 4(a): circles show token/token, squares byte/byte and diamonds byte/token, with the readout named first. Separate rows identify the settings. Comparing complete procedures gives the difference between the two matched points. Each arrow connects one matched point to the crossed setting and represents a change in one rule with the other fixed. Points are means over three equally weighted tasks with 95% paired percentile bootstrap intervals (2000 replicates); arrow endpoints give the corresponding risk estimates. Huginn observes steps 16–128 and Ouro passes 2–4. The convention pair in **(a)** is byte mean against sum, and in **(b)** byte mean against token mean.

A worked evaluation example.

On the new Ouro questions, the two matched procedures differed by only 0.67 pp. An evaluator who compared only these procedures might conclude that the token and byte margins are interchangeable. The crossed setting tests that conclusion directly: with the byte readout fixed, selecting by the token margin raised risk by 12.00 pp at the same release

quota. The matched comparison itself is valid; the unsupported step is reading its small difference as evidence about the margins under a fixed answer rule. The example concerns measured revision risk and shows neither an accuracy loss nor a changed budget decision.

Equation (3) splits the complete-procedure comparison at fixed release into a margin leg, which swaps the margin at a fixed byte-mean readout, and a readout leg, which swaps the readout at a fixed token-mean margin. The margin leg ranged over 9.33 to 12.00 percentage points (pp) and the readout leg over 9.67 to 13.00 pp. Both legs had intervals excluding zero in all 3 cohorts. Their difference, the contrast between complete procedures, ranged over 0.33 to 3.00 pp. Its interval included zero in every cohort; these intervals do not establish equivalence between complete procedures. The arithmetic consistency checks are retained in Supplementary Section S16.4.

Equation (3) does not require two large opposing changes: if both readouts assigned identical revision labels to the questions selected by the token margin, the readout leg would be zero.

Averaging both readout directions over six cells, as the original Ouro protocol's primary estimand did, gives 10.00 pp ([7.16, 12.50]) for Huginn and 11.00 pp ([7.50, 13.67]) on the same historical data. The second crossed setting contributes a contrast of the same sign; this average is a distinct estimand computed from the same outputs.

Table 5 exposes the task-level fixed-byte-readout effects behind the averages. Each task has 200 questions per cohort, and the historical models use the same questions. These rows describe heterogeneity; they are not nine independent replications. The revision's Bonferroni assessment keeps all six headline margin/readout legs above zero for family H ($m = 6$), with adjusted lower bounds ranging from 3.67 to 5.67 pp. For family P ($m = 3$), the historical Ouro primary is 11.00 pp [7.00, 14.50], the new Ouro C effect is -11.33 pp $[-15.67, -4.67]$, and Q is -1.67 pp $[-3.33, 2.33]$. These approximate adjusted intervals do not establish Q equivalence or change the post hoc status of the separate legs.

Table 5. Task-level margin leg at 50% release with the byte readout fixed (any-time revision, pp). Brackets give pointwise 95% intervals from the revision's 20,000 draws. Historical questions are shared between models. Full four-cell risks and endpoint results are supplied with Supplementary Section S16.

Task	Huginn Historical	Ouro Historical	Ouro New Questions
ARC-Challenge	6.00 [0.00, 11.00]	10.00 [3.00, 17.00]	7.00 [−2.00, 13.00]
CommonsenseQA	19.00 [11.00, 26.00]	13.00 [4.00, 22.00]	11.00 [4.00, 20.00]
OpenBookQA	3.00 [−1.00, 9.00]	7.00 [−2.00, 16.00]	18.00 [7.00, 24.00]

Within each length unit, the sum and mean rules were endpoints of a broader pattern. These grids examine a different parameter space from the byte/token comparison: each holds the length unit fixed and moves from sum to mean. On the exponent grid of Equation (6), with the readout fixed at exponent α , the lowest risks occurred near the matching margin exponent (Figure 2). Cells at least 0.5 apart carried 8.25–11.54 pp more risk than the diagonal in all 6 grids across cohorts and units, with every interval excluding zero and a smallest lower bound of 6.39 pp. The mean risk at each distance increased with distance in every grid (Spearman 0.973–1.000). The empirical minimiser on the tested grid moved with the readout: the chosen exponent lay within 0.2 of the readout exponent in at least 10 of 11 rows of every grid and took 6 to 8 distinct values. A separate check retaining all tied minima found an empty intersection across readouts in every grid. Thus no tested margin exponent attained the empirical minimum for every readout, even allowing ties. The diagonal cell exceeded its row minimum by at most 3.00 pp. All three registered criteria passed.

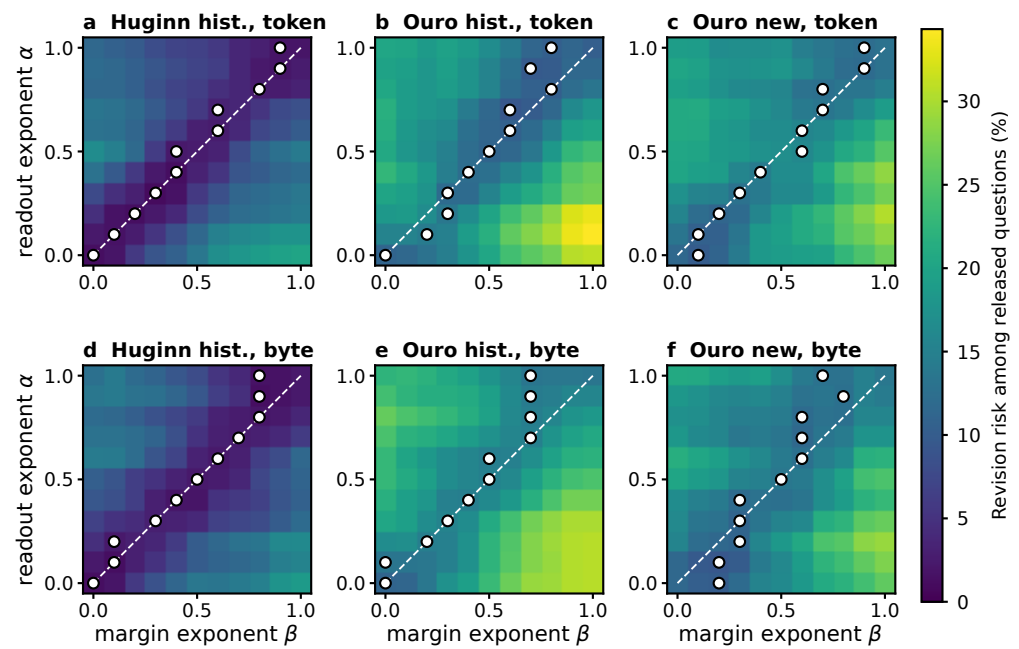


Figure 2. Revision risk on the exponent grid. Each panel shows $r(\alpha, \beta)$ at 50% release for one cohort and one length unit: rows give the readout exponent α and columns the margin exponent β of Equation (6), and colour gives the revision risk among released questions, averaged over the three tasks. White points mark the smallest margin exponent within 10^{-9} percentage points of the minimum unrounded risk in each row; the dashed line is the matched diagonal. The token-grid corners cross sum and token mean; the byte-grid corners cross sum and byte mean. Cloze byte and character lengths coincide in these cohorts, so the latter correspond to the sum/character pair in Table 6. The byte/token crossed cells of Table 4 are outside both grids. The corner results were known before the interior-grid criteria were fixed. Minima refer only to the tested grid points. (a–c) token unit; (d–f) byte unit.

Table 6. Revision risks (%) for each convention pair (v, w) at 50% release. Matched settings use the same rule for margin and readout; crossed settings mix them. The last column is the smallest crossed risk minus the largest matched risk. The criterion was fixed before this calculation on existing data (A in Table 3). Cohorts and task averaging are as in Table 4; intervals are pointwise and unadjusted.

Cohort	Pair (v, w)	$r(v, v)$	$r(w, w)$	$r(v, w)$	$r(w, v)$	Crossed – Matched
Huginn, earlier	byte/token	2.33	2.00	11.67	12.67	9.33 [5.00, 11.33]
	sum/character	1.33	2.33	18.67	12.67	10.33 [6.33, 14.00]
	sum/token	1.33	2.00	20.00	11.67	9.67 [6.00, 13.33]
Ouro, earlier	byte/token	13.33	10.33	23.33	22.33	9.00 [3.33, 11.67]
	sum/character	11.00	13.33	30.67	22.33	9.00 [3.32, 13.67]
	sum/token	11.00	10.33	31.00	20.33	9.33 [3.00, 13.67]
Ouro, new	byte/token	13.67	13.00	25.67	19.67	6.00 [0.67, 9.67]
	sum/character	11.67	13.67	26.00	20.83	7.17 [1.67, 11.33]
	sum/token	11.67	13.00	26.67	19.17	6.17 [1.33, 10.33]

Two limits are visible in the figure. The diagonal itself was not flat, and beyond a distance of 0.8 the rise levelled off in the Ouro cohorts; the difference between the farthest and the middle distance included zero in 1 of the 6 grids (Supplementary Section S13).

3.3. Higher Risk in Crossed Settings Extended to Other Conventions and Exit Signals

The ordering also held for the other convention pairs available in the data. The frozen scores allow three distinct pairs: byte/token, sum/character and sum/token. Both

matched settings sat below both crossed settings in 9 of 9 combinations of cohort and pair (Table 6). The smallest crossed risk exceeded the largest matched risk by 6.00–10.33 pp, every interval excluded zero, and the ordering held in at least 98.5% of bootstrap replicates in every cell. Cloze character and byte divisors coincide for the scored continuations in these three cohorts, including their leading spaces, so the two remaining pairs duplicate these comparisons. H1 character lengths use a different text boundary (Table 2).

The crossed setting is also measurable with the signals adaptive-depth models use to stop. Each model’s own exit signal released the 50% of questions it marked as most converged at the anchor, with revision still defined by the byte mean. That release carried 10.00 pp ([5.67, 13.67]) more risk than the matched byte setting for Huginn’s step-to-step divergence, and 9.33 ([4.33, 14.67]) and 7.00 ([2.00, 12.00]) more for Ouro’s learned gate (Table 7). All 14 contrasts across signals and cohorts were positive with intervals excluding zero. The questions selected by the exit signal had higher risk than the matched set in at least 99.5% of replicates. The models’ own signals therefore reproduced the excess risk over the matched setting. Their risk did not consistently reach or exceed that of the crossed setting using the token margin, so the second of the two criteria registered for this comparison failed. The differences were 0.67 pp ([−3.33, 5.00]), −0.67 ([−5.67, 4.67]) and −5.00 ([−9.00, 1.33]). Ouro’s hidden-state displacement, the alternative heuristic its report tests, was the worst selector, 18.67 pp above the matched setting. Huginn’s default thresholds released at most 2 questions per task at the anchor, so only quota release is comparable.

Table 7. The models’ exit signals at 50% release with the byte readout fixed. Each signal selects the questions it marks as most converged at the anchor. Columns give revision risk, excess over matched byte, and difference from token crossed. Criteria were fixed before this comparison on existing data (A in Table 3). Cohorts, task averaging and bootstrap draws are as in Table 4; intervals are pointwise and unadjusted. KL: Kullback–Leibler divergence.

Cohort	Exit Signal	Risk (%)	Minus Matched (pp)	Minus Token Crossed (pp)
Huginn, earlier	KL between steps	12.33	10.00 [5.67, 13.67]	0.67 [−3.33, 5.00]
	latent change	10.00	7.67 [4.67, 11.67]	−1.67 [−4.62, 3.00]
	min-p KL	12.00	9.67 [5.99, 13.67]	0.33 [−3.22, 5.00]
	latent cosine distance	13.66	11.32 [8.50, 14.30]	1.99 [−0.68, 5.68]
	argmax stability	13.77	11.44 [7.60, 15.21]	2.11 [−1.45, 6.75]
	entropy change	15.33	13.00 [9.00, 17.00]	3.67 [0.00, 8.33]
Ouro, earlier	gate, cumulative	22.67	9.33 [4.33, 14.67]	−0.67 [−5.67, 4.67]
	gate, prompt mean	21.67	8.33 [2.67, 13.33]	−1.67 [−7.00, 3.33]
	gate, instantaneous	22.67	9.33 [4.33, 14.60]	−0.67 [−5.67, 4.75]
	hidden-state displacement	32.00	18.67 [13.67, 23.33]	8.67 [3.33, 14.00]
Ouro, new	gate, cumulative	20.67	7.00 [2.00, 12.00]	−5.00 [−9.00, 1.33]
	gate, instantaneous	21.33	7.67 [2.00, 12.44]	−4.33 [−8.81, 1.56]
	gate, prompt mean	22.00	8.33 [3.00, 13.33]	−3.67 [−8.33, 2.34]
	hidden-state displacement	32.33	18.67 [13.00, 23.00]	6.67 [1.67, 12.34]

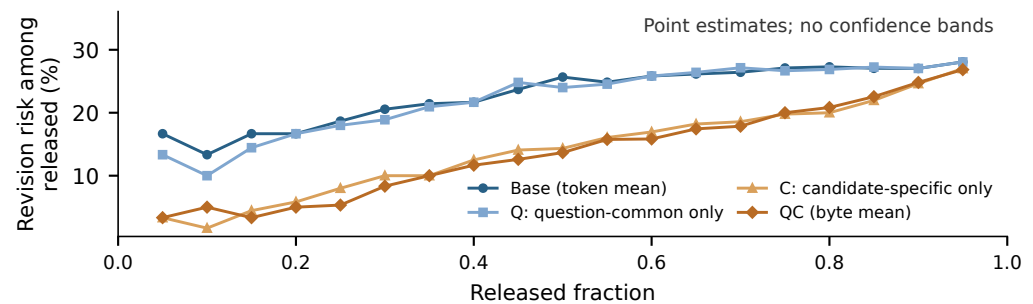
The additional exit-signal ranking check found no reversal with both contrast intervals excluding zero. It therefore provides no evidence that changing the readout changes the preferred exit signal; the counts and criterion are retained in Supplementary Section S16.

3.4. Rescaling Individual Candidates Gave the Larger Estimated Risk Reduction

With the byte answer rule fixed, we compare question-wide scaling (Q), candidate-specific scaling (C), and their combination (QC), as defined in Section 2.3. On the new Ouro cohort, the mean revision risks for base, Q, C and QC were 25.67%, 24.00%, 14.33% and

13.67%. Relative to base, Q changed risk by -1.67 pp ($[-3.00, 1.67]$) and C by -11.33 pp ($[-14.67, -5.67]$). The total QC contrast was -12.00 pp ($[-15.33, -6.33]$); the interaction $QC - C - Q + \text{base}$ was 1.00 pp ($[-3.00, 2.67]$). The candidate-specific operation therefore gave the larger reduction in point estimate in this cohort; the four cells are shown across release fractions in Figure 3a. The total equals the negative of the same cohort's margin leg by construction and is not independent evidence.

a Four scoring cells, Ouro new questions



b Margin leg, all three cohorts

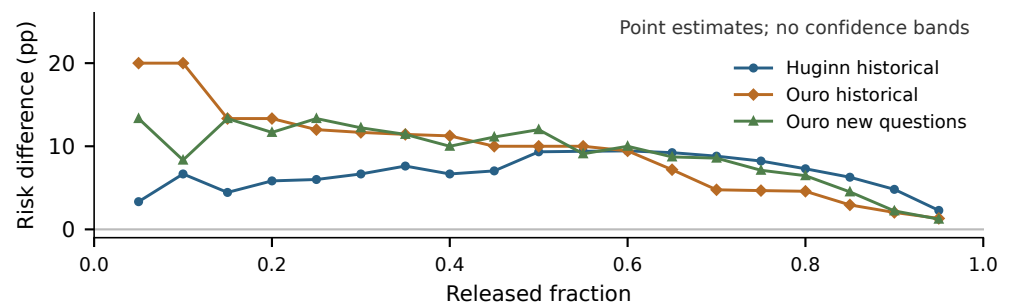


Figure 3. Scoring-rule effects across the release fraction, swept over 19 points spanning 0.05–0.95. (a) Revision risk among released questions for the four scoring cells of Section 3.4 on the new Ouro cohort, with the readout held at the byte mean. (b) The risk difference from changing the margin for each cohort. Both panels average the three tasks equally. At 50% release, panel (a) reproduces the four cell risks reported in Section 3.4 and panel (b) the margin legs of Figure 1b. Curves are point estimates without intervals, and the sweep reuses the same questions at every release fraction, so neighbouring points are strongly dependent. The scan is descriptive: it calibrates no policy and adds no replication.

The Q interval is compatible with both zero and nonzero effects, leaving equivalence unestablished. Candidate-specific rescaling can change both the candidate leader and numerical gaps, including the ranking of questions. The C contrast therefore locates an arithmetic operation associated with the measured change without identifying leader changes as the unique cause. Earlier factor analyses on historical Huginn and Ouro outputs remain exploratory support; the locally prospective new-question factor study used Ouro only.

3.5. Release Fraction, Observation Window and Task Accuracy Limit the Interpretation

The 50% release point is a convention, so the margin leg was recomputed at 19 release fractions spanning 0.05–0.95 (Figure 3b). It stayed positive at every one of them in all three cohorts, ranging over 2.28 to 9.44 pp for Huginn historical, 1.32 to 20.00 pp for Ouro historical and 1.23 to 13.33 pp for the new Ouro cohort. The sign was consistent across the tested operating points, while the magnitude varied. All three curves shrink toward zero as the released fraction approaches the whole cohort, which they must, since at full release the two margins select the same set. Away from that limit the magnitude is not monotone:

both Ouro cohorts are largest at the smallest release fractions, whereas the Huginn curve peaks near the middle of the range. No single operating point is representative. The scan reuses the same questions at every point, so the curves are not independent observations, and no calibrated threshold for individual questions was established.

The revision outcome also matters. On all 600 Huginn questions, 36 of 85 byte-mean revisions returned to the anchor answer by the horizon: 42.4% of revisions, or 6.0% of questions. The corresponding byte-mean return fractions among revisions were 6.8% and 7.6% in historical and new Ouro. Under token mean they were 42.4%, 9.3% and 10.1%. These whole-cohort counts are distinct from risk among releases. At the headline windows, replacing any-time revision by final-depth disagreement left fixed-byte-readout margin effects of 4.67 pp [2.00, 7.67], 9.33 pp [4.67, 14.00] and 10.33 pp [5.00, 14.00]. Thus the effect persists, but Huginn's magnitude roughly halves. Both crossed risks still exceeded both matched risks in point estimate; the new Ouro gap interval included zero (4.33 pp [−0.33, 8.00]). Supplementary Section S16 gives all four risks under both outcomes.

Across the 13 Huginn windows, the any-time margin leg ranged from 2.33 to 16.00 pp. It ranged from 3.67 to 10.00 pp across the six historical Ouro windows and from 7.33 to 12.00 pp across the six new Ouro windows. A roughly 10 pp effect is therefore specific to the headline setting. At Huginn steps 64–128, the final-depth margin leg was only 0.33 pp [0.00, 1.00]. The earlier Huginn-at-Ouro-window check also retains its original verdict: both criteria failed, one because the readout-leg interval touched zero, the other because the overlap interval included chance. Further window checks do not reverse those failures. Equal step indices do not make the models' computation costs equal.

The lower revision risk did not establish an improvement in task accuracy. For the six-cell means, hybrid accuracy in crossed settings minus that in matched settings was 0.08 pp ([−0.58, 0.83]) for Huginn and −1.00 pp ([−2.25, 0.33]) for Ouro; both intervals contain zero. These intervals use whole-cohort accuracy, whereas revision risk conditions on release. Early revisions can include both corrections and degradations, which is consistent with the absence of an established accuracy gain.

The historical first-crossing replay changed none of the six program–task budget classifications after thresholds were refitted. Its details remain in Supplementary Section S14, with a summary in Supplementary Section S16.4. Together with the accuracy and exit-ranking results, this limits the demonstrated practical consequence to the selection and interpretation of revision risk. We have not established an improved stopping decision or a computation saving. Equal release fractions alone do not account for the cost of scoring candidate answers.

3.6. The Two-Point Test Bounds the Dependence Within a Named Length Unit

The two-point test of Section 2.4 identifies questions whose answer is invariant within a specified normalization family. Applied to the token unit on 1800 cases across the three model cohorts, it certified 910 of them as invariant over every exponent between the sum and the token mean, or 49.2–53.3% per cohort. Direct checks found 0 counterexamples: a 201-point grid in α for the one-unit statement, and the four vertices plus a 21×21 interior grid of the token/character rectangle for Corollary 1.

The length unit is essential. Of the 910 token-certified cases, 267 still changed winner between token and character means. Exactly equal token counts guarantee invariance for only 310 cases; the two-point test uses the actual scores and covers more. This is a fixed-depth arithmetic guarantee, not a guarantee of stable future answers or an unchanged released set. The full unit comparisons and archived-evaluation coverage are retained in Supplementary Sections S13, S15 and S16.

The certificate tests whether a named normalization family can change the winner. Ranking the remaining questions or choosing when to stop would require additional predictive value. Four exploratory tests of those uses were negative and are reported in Supplementary Section S13 and Supplementary Data S1.

4. Discussion

The main finding is the size of the dependence between question selection and answer definition. At 50% release, the byte and token complete procedures differed by only 0.33 to 3.00 pp. Holding one rule fixed exposed changes of roughly 10 pp. These larger changes reveal dependence that the complete-procedure comparison conceals. Such a comparison correctly describes the two procedures, but it cannot establish that their selectors are interchangeable. This extends established findings on length bias and prediction churn [1,2,9] to selected revision events along recurrent trajectories.

The outcome definition changes the size of the effect. Many Huginn revisions were transient, and final-depth disagreement reduced its fixed-readout margin effect from 9.33 to 4.67 pp. The corresponding Ouro effects remained above 9 pp. Later Huginn anchors gave much smaller differences. These results make the window and the revision event part of the quantity being measured, not incidental evaluation settings. The sum-to-mean grids provide a separate view: low-risk margins generally followed the readout within each length unit. Their minima are empirical minima among tested points, not validated optimal exponents or stopping rules.

For benchmark designers, the practical rule is to fix the answer rule and the revision event before comparing selectors. Use the corresponding matched margin as a reference, then evaluate alternative margins against that same target. If more than one answer rule is plausible, report the crossed risk table as well as the matched comparison. This reveals whether a selector's apparent advantage depends on how an answer is defined. Matching is a useful reference, not a guarantee of the best selector. Report the release fraction, anchor, horizon and paired question-level outcomes so that others can reproduce the comparison.

The interaction affects both the reported risk and which questions a margin selects for early release. Lower revision risk did not establish a hybrid accuracy gain, the historical budget classifications were unchanged, and the exit-signal ranking check found no reversal supported by intervals excluding zero. Equal release fractions also do not imply equal computation. Native exit signals use prompt states; candidate margins require scoring the answer choices. A policy comparison would need task accuracy, the cost of obtaining each signal, total computation or latency, and thresholds assessed on data independent of their selection. Consistency and task performance remain distinct objectives [18,19,22].

The evidence has a limited scope: two models, three filtered benchmarks, cloze scoring, and 200 questions per task and cohort. The historical models share questions, and repeated analyses of stored outputs do not add independent replications. The new OpenBookQA cohort uses validation rather than the historical test split. "New questions" means newly collected for this study; their pretraining exposure is unknown. OLMES recommends PMI scoring for these tasks, but the required unconditional likelihoods were unavailable. The principal results therefore concern byte and token cloze scores, with sum-to-mean grids as supporting analyses.

Transfer to dynamic latent computation is plausible when a system still scores the same candidate answers at successive depths. In that setting, a selector and the answer rule can be crossed in the same way, although the effect size must be measured again. Free-form generation changes the problem. Decoding can alter both the generated strings and their lengths, and revision may mean token disagreement, a changed final answer or a semantic change. A fixed candidate set and its top-two margin may not exist. We therefore

expect the need to specify selection and outcome separately to transfer more broadly than the numerical effects reported here. We have not tested a generation or exact-match task, and make no quantitative transfer claim.

The historical H1 harness comparison needs an additional qualification. Common-senseQA uses unnormalized answer-letter likelihoods in H1, whereas ARC and Open-BookQA use character-normalized answer-text likelihoods. At step 32 on Common-senseQA, H1 achieved 20.3% accuracy on 256 source questions and 23.5% on 200 evaluation questions. It selected B on 200/256 and 163/200 questions, respectively. This bias can make an inaccurate answer look stable. The cause remains unresolved despite checks of the template, option order, gold mapping and likelihood-to-answer mapping. H1 and the cloze program also differed in prompting, batching and initialization, so their historical comparison cannot isolate a readout effect. H1 outputs are absent from the principal crossed cloze analysis. Supplementary Sections S4 and S8 retain the historical results with these limitations.

The two-point certificate addresses a narrower question than future stability. Proposition 1 checks the winner at one observed depth over a specified normalization interval and length unit. It uses stored scores at the normalization endpoints and needs no additional model forward pass. In contrast, Supplementary Theorem S1 concerns decoded outputs over future iterations of a native dynamical system. Its regional contraction and phase-image conditions were not closed by the available bounds. We therefore have an exact arithmetic test for normalization invariance, but no native finality certificate for these model trajectories. The archived-evaluation applications and unsuccessful predictive uses of the arithmetic test remain in Supplementary Sections S13, S15 and S16.4 [31].

5. Conclusions

Similar revision risks from two complete procedures can conceal a substantial dependence between selection and answer definition. In our headline comparisons, changing either the margin or the readout moved risk by roughly 10 pp, while changing both moved it by only 0.33 to 3.00 pp. The size of this dependence varied with the revision outcome and observation window.

The evidence covers Huginn and Ouro, cloze scoring on three filtered benchmarks, 200 questions per task and cohort, and a headline release fraction of 50%. The historical cohorts share questions. Within this scope, the results support a reusable evaluation practice: compare selectors under a fixed answer rule, then cross the rules to show how much the conclusion depends on the revision target. Transfer to other models or generation tasks, and gains in accuracy or stopping utility, require separate evaluation.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app16199857/s1>. Supplementary Information: Sections S1–S16, Figures S1–S8, Tables S1–S28, Theorem S1 and Corollary S1, covering scoring methods, supporting studies with their original evidence status, and revision outcome, window sensitivity and multiplicity analyses. References cited in the Supplementary Information are included in the main reference list [2,22–24,32–36]. Supplementary Data S1 (Supplementary_Data_S1.zip): stored numerical trajectories, analysis code, source data, protocols and supporting results, including the revision analyses and the `replay_all.py` entry point.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, writing (original draft preparation) and writing (review and editing), R.L.; supervision and writing (review and editing), Y.R. and Z.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China, grant number 2018AAA0100204.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and code supporting this study are supplied in Supplementary Data S1. The archive contains stored candidate scores, numerical length and label metadata, question identifiers, source hashes, analysis code and figure source data. The offline replay reproduces the eleven main effect and accuracy point estimates, three paired revision-label tables, historical first-crossing results and the six exponent-grid risk tables. A numerical port of the original bootstrap procedure regenerates the risks and contrasts in Table 4 and the core grid summaries behind Figure 2, including their confidence intervals, from the stored numerical inputs. It also checks all tied grid minima. Other supporting intervals and historical analyses remain outside this replay’s execution scope and are supplied as original exports with analysis code. Benchmark question and candidate text are available from the original sources listed in the archive; the supplied numerical data suffice for offline replay. Model weights are obtained from their original repositories. Transformed release metadata and original collection hashes are identified separately. The revised archive also reproduces the endpoint, transient-revision, window and multiplicity analyses and their published table entries through `replay_all.py`. The replay checks stored outputs; it repeats no model inference and provides no external preregistration.

Acknowledgments: The authors used OpenAI Codex (GPT-5.6 and GPT-6 model families) and Anthropic Claude Code (Claude Opus 5 and Claude Fable 5 model families) for research assistance, code development, mathematical and statistical checks, literature searches, and manuscript preparation. DeepSeek (deepseek-flash V4) assisted manuscript editing. Benchmark data were not generated by these assistants; numerical results were computed from the specified model outputs and analysis procedures. AI-assisted checks do not constitute external validation or independent peer review. The authors reviewed and edited the analyses, code, arguments, references and text, and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gu, Y.; Tafjord, O.; Kuehl, B.; Haddad, D.; Dodge, J.; Hajishirzi, H. OLMES: A Standard for Language Model Evaluations. In *Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 5020–5048. <https://doi.org/10.18653/v1/2025.findings-naacl.282>.
2. Oostermeijer, K. Accuracy and Normalized Accuracy under Length Bias: Analysis, Guidelines, and a Bayesian Alternative. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, Seoul, Republic of Korea, 6–11 July 2026. <https://doi.org/10.48550/arXiv.2607.12767>.
3. Holtzman, A.; West, P.; Shwartz, V.; Choi, Y.; Zettlemoyer, L. Surface Form Competition: Why the Highest Probability Answer Isn’t Always Right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 7038–7051. <https://doi.org/10.18653/v1/2021.emnlp-main.564>.
4. Robinson, J.; Rytting, C.M.; Wingate, D. Leveraging Large Language Models for Multiple Choice Question Answering. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, 1–5 May 2023. <https://doi.org/10.48550/arXiv.2210.12353>.
5. Alzahrani, N.; Alyahya, H.; Alnumay, Y.; Alrashed, S.; Alsubaie, S.; Almushayqih, Y.; Mirza, F.A.; Al-Twairesh, N.M.A.; Alowisheq, A.; Bari, M.S.; et al. When Benchmarks are Targets: Revealing the Sensitivity of Large Language Model Leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 13787–13805. <https://doi.org/10.18653/v1/2024.acl-long.744>.
6. Biderman, S.; Schoelkopf, H.; Sutawika, L.; Gao, L.; Tow, J.; Abbasi, B.; Aji, A.F.; Ammanamanchi, P.S.; Black, S.; Clive, J.; et al. Lessons from the Trenches on Reproducible Evaluation of Language Models. *arXiv* **2024**. <https://doi.org/10.48550/arXiv.2405.14782>.
7. Michaelov, J.A. `acc_norm`: Character vs Byte Length. EleutherAI/lm-Evaluation-Harness Issue #3278. Opened 2 September 2025; Confirmed by a Repository Collaborator on 10 September 2025; Open as of 11 September 2026. Available online: <https://github.com/EleutherAI/lm-evaluation-harness/issues/3278> (accessed on 11 September 2026).

8. EleutherAI. LM Evaluation Harness: Metrics. Distinguishes Character-Normalized acc_norm and Byte-Normalized acc_bytes. Available online: <https://lm-evaluation-harness.readthedocs.io/api/metrics/> (accessed on 4 September 2026).
9. Milani Fard, M.; Cormier, Q.; Canini, K.; Gupta, M. Launch and Iterate: Reducing Prediction Churn. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29.
10. Livingston, S.A.; Lewis, C. Estimating the Consistency and Accuracy of Classifications Based on Test Scores. *J. Educ. Meas.* **1995**, *32*, 179–197. <https://doi.org/10.1111/j.1745-3984.1995.tb00462.x>.
11. Bland, J.M.; Altman, D.G. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *Lancet* **1986**, *327*, 307–310. [https://doi.org/10.1016/S0140-6736\(86\)90837-8](https://doi.org/10.1016/S0140-6736(86)90837-8).
12. Kim, H.; Kang, P. Same Answer, Different Confidence: Protocol Sensitivity in LLM Confidence Calibration. Version 3, 7 August 2026. *arXiv* **2026**. <https://doi.org/10.48550/arXiv.2605.27752>.
13. Graves, A. Adaptive Computation Time for Recurrent Neural Networks. *arXiv* **2016**. <https://doi.org/10.48550/arXiv.1603.08983>.
14. Dehghani, M.; Gouws, S.; Vinyals, O.; Uszkoreit, J.; Kaiser, Ł. Universal Transformers. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019. <https://doi.org/10.48550/arXiv.1807.03819>.
15. Banino, A.; Balaguer, J.; Blundell, C. PonderNet: Learning to Ponder. In Proceedings of the 8th ICML Workshop on Automated Machine Learning (AutoML), Virtual, 23 July 2021. <https://doi.org/10.48550/arXiv.2107.05407>.
16. Zhou, W.; Xu, C.; Ge, T.; McAuley, J.; Xu, K.; Wei, F. BERT Loses Patience: Fast and Robust Inference with Early Exit. In Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS 2020), Online, 6–12 December 2020.
17. Schuster, T.; Fisch, A.; Jaakkola, T.; Barzilay, R. Consistent Accelerated Inference via Confident Adaptive Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 4962–4979. <https://doi.org/10.18653/v1/2021.emnlp-main.406>.
18. Schuster, T.; Fisch, A.; Gupta, J.; Dehghani, M.; Bahri, D.; Tran, V.; Tay, Y.; Metzler, D. Confident Adaptive Language Modeling. In Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS 2022), New Orleans, LA, USA, 28 November–9 December 2022.
19. Jazbec, M.; Timans, A.; Hadži Veljković, T.; Sakmann, K.; Zhang, D.; Næsseth, C.A.; Nalisnick, E. Fast yet Safe: Early-Exiting with Risk Control. In Proceedings of the Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Vancouver, BC, Canada, 10–15 December 2024. <https://doi.org/10.52202/079017-4124>.
20. Geifman, Y.; El-Yaniv, R. Selective Classification for Deep Neural Networks. In *Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017)*; Curran Associates, Inc.: Red Hook, NY, USA, 2017. <https://doi.org/10.48550/arXiv.1705.08500>.
21. Jazbec, M.; Forré, P.; Mandt, S.; Zhang, D.; Nalisnick, E. Early-Exit Neural Networks with Nested Prediction Sets. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, 15–19 July 2024*; PMLR: Cambridge, MA, USA, 2024; Volume 244, pp. 1780–1796.
22. Popescu, A.C.; Sáez de Ocáriz Borde, H.; Liò, P. Adaptive Depth in Looped Transformers: Diagnosing Learned Halting Gates and Trajectory Readouts. Version 1. *arXiv* **2026**. <https://doi.org/10.48550/arXiv.2607.20519>.
23. Geiping, J.; McLeish, S.; Jain, N.; Kirchenbauer, J.; Singh, S.; Bartoldson, B.R.; Kailkhura, B.; Bhatele, A.; Goldstein, T. Scaling up Test-Time Compute with Latent Reasoning: A Recurrent Depth Approach. *arXiv* **2025**. <https://doi.org/10.48550/arXiv.2502.05171>.
24. Zhu, R.J.; Wang, Z.; Hua, K.; Zhang, T.; Li, Z.; Que, H.; Wei, B.; Wen, Z.; Yin, F.; Xing, H.; et al. Scaling Latent Reasoning via Looped Language Models. *arXiv* **2025**. <https://doi.org/10.48550/arXiv.2510.25741>.
25. Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; Tafjord, O. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv* **2018**. <https://doi.org/10.48550/arXiv.1803.05457>.
26. Talmor, A.; Herzig, J.; Lourie, N.; Berant, J. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4149–4158. <https://doi.org/10.18653/v1/N19-1421>.
27. Mihaylov, T.; Clark, P.; Khot, T.; Sabharwal, A. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 2381–2391. <https://doi.org/10.18653/v1/D18-1260>.
28. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. <https://doi.org/10.1177/001316446002000104>.
29. Feinstein, A.R.; Cicchetti, D.V. High Agreement but Low Kappa: I. The Problems of Two Paradoxes. *J. Clin. Epidemiol.* **1990**, *43*, 543–549. [https://doi.org/10.1016/0895-4356\(90\)90158-L](https://doi.org/10.1016/0895-4356(90)90158-L).
30. Adler, I.; Monteiro, R.D.C. A Geometric View of Parametric Linear Programming. *Algorithmica* **1992**, *8*, 161–176. <https://doi.org/10.1007/BF01758841>.

31. Hugging Face. Open LLM Leaderboard (Archived v1): Per-Model Evaluation Records. Hugging Face Hub, open-llm-leaderboard-old. Available online: <https://huggingface.co/collections/OpenEvals/archived-open-llm-leaderboard-2023-2024> (accessed on 8 September 2026).
32. EleutherAI. LM Evaluation Harness: Task Definitions and Model Adapter. Version 0.4.7, 2024. Available online: <https://github.com/EleutherAI/lm-evaluation-harness/tree/4c26a9c176ecfb40b369503ce211e356bb800fdc> (accessed on 4 September 2026).
33. Huang, B.; Geng, Z.; Kolter, J.Z. Equilibrium Reasoners: Learning Attractors Enables Scalable Reasoning. In Proceedings of the 43rd International Conference on Machine Learning (ICML), Seoul, Republic of Korea, 6–11 July 2026. <https://doi.org/10.48550/arXiv.2605.21488>.
34. Jolicoeur-Martineau, A. Less Is More: Recursive Reasoning with Tiny Networks. *arXiv* **2025**. <https://doi.org/10.48550/arXiv.2510.04871>.
35. Maia Polo, F.; Weber, L.; Choshen, L.; Sun, Y.; Xu, G.; Yurochkin, M. tinyBenchmarks: Evaluating LLMs with Fewer Examples. In *Proceedings of the 41st International Conference on Machine Learning*; PMLR, 2024; Volume 235, pp. 34303–34326. Available online: <https://proceedings.mlr.press/v235/maia-polo24a.html> (accessed on 4 September 2026).
36. Wang, G.; Li, J.; Sun, Y.; Chen, X.; Liu, C.; Wu, Y.; Lu, M.; Song, S.; Abbasi Yadkori, Y. Hierarchical Reasoning Model. *arXiv* **2025**, <https://doi.org/10.48550/arXiv.2506.21734>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.