OFF-MENU PREFERENCES DIAGNOSE MENU SENSITIVITY IN FORCED-CHOICE EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Forced-choice benchmarks score a language model only on the answers they list (the *menu*), yet a model can score an unlisted answer above every listed one. We ask whether such off-menu preferences indicate sensitivity to the menu or an incorrect answer. A gold-free off-menu ratio compares the best-scoring candidate in a specified excluded set with the top listed answer; a controlled swap replaces the weakest distractor with that candidate while retaining the gold answer. Across four open-weight checkpoints, controlled cloze tasks, and four established benchmarks, the menu can matter more than the model: on 941 single-token CommonsenseQA items, the swap lowers Qwen2.5-1.5B's accuracy from 66.0% to 21.6%, a 44.4-point drop, more than twice the largest accuracy gap between checkpoints (20.3 points). Because the inserted candidate is never the gold answer, the drop equals the proportion of items on which it displaces a correct answer; displacement rate times overall accuracy tracks drops of 3 to 44 points across three checkpoints and four benchmarks (correlation 0.998). At equal coverage, the margin between the top two listed answers is the stronger selector of correct answers (81.1% vs. 68.4% on CommonsenseQA; likewise on ARC and OpenBookQA). The ratio instead flags risk under the altered menu: used in labelled calibration at 50% coverage, it lowers the proportion of Qwen2.5-1.5B's samples from that menu that miss the gold answer from .652 to .610. Evaluations should report the menu, the excluded-candidate search, and the calibration target alongside accuracy.

1 INTRODUCTION

Multiple-choice (forced-choice) benchmarks are a primary way to compare language models (Brown et al., 2020; Hendrycks et al., 2021; Robinson & Wingate, 2023). They make an open prediction problem tractable by restricting the answer set, which we call the *menu*. This restriction is itself part of the measurement. A model may reliably choose one listed answer and yet assign a higher score to a continuation that the menu excludes (Mensa et al., 2026; Wiegreffe et al., 2025; Abdallah et al., 2026). Figure 1(a) illustrates this pattern with schematic scores: the model prefers *cupboard* among the listed options but scores the unlisted *kitchen* higher still. Such an *off-menu preference* raises two questions. Does it reveal a problem with the menu or an incorrect answer? How much of a benchmark comparison reflects the menu rather than the model?

Answering these questions requires separating three targets. *Within-menu confidence* concerns the relative scores of the listed candidates. *Menu sensitivity* concerns how the response changes when the candidate set changes. *Answer correctness* concerns a benchmark label or a semantic judgment. Standard practice stays inside the menu: benchmark accuracy and sampling-based aggregation such as self-consistency (Wang et al., 2023) both work with the listed answers, so an omission becomes visible only when candidates beyond the menu are scored. Robustness studies vary how the listed options are presented, such as their order and identifiers (Pezeshkpour & Hruschka, 2024; Zheng et al., 2024), and separate choice-driven from question-driven scores (Cho et al., 2026); in these studies the set of listed candidates stays fixed.

We measure menu sensitivity directly with two tools (Figure 1(b)). The *off-menu ratio* compares two scores on a common scale: that of the best-scoring *excluded candidate* (a candidate absent from the menu) and that of the *listed leader* (the highest-scoring listed answer). It uses no gold label. The *swap* keeps the question and the gold answer and replaces only the lowest-scoring distractor with

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

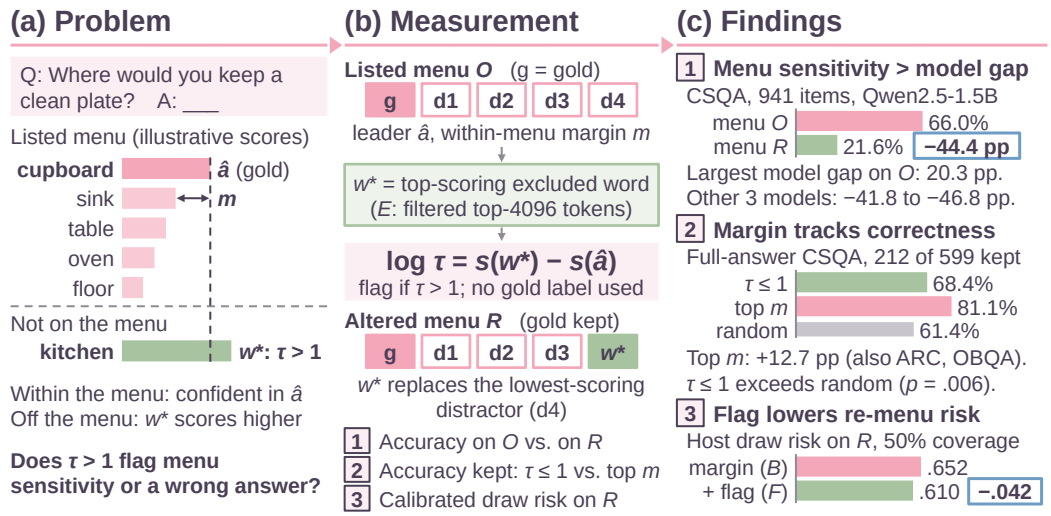

Figure 1: **Off-menu preferences diagnose menu sensitivity, a target distinct from answer correctness.** (a) An illustrative item: the model is confident among the listed options, yet the excluded candidate *kitchen* scores above the listed leader $\hat{a}$ ($\tau > 1$); m is the within-menu margin. (b) The gold-free off-menu ratio τ compares the best-scoring candidate w^* in a specified excluded set $\mathcal{E}$ with $\hat{a}$; the swap replaces the lowest-scoring distractor by w^* and keeps the gold answer. (c) Results (CSQA: CommonsenseQA; OBQA: OpenBookQA). (1) On the 941-item single-token CSQA slice, the swap lowers the host checkpoint's (Qwen2.5-1.5B) accuracy from 66.0% to 21.6%, more than twice the largest original-menu gap between the four checkpoints (20.3 percentage points, pp). (2) On a separate full-answer CSQA evaluation half, the 212 items with the largest m are more often correct than the 212 ratio-clean items (81.1% vs. 68.4%; likewise on ARC and OBQA), which exceed random subsets of equal size (61.4%, upper-tail $p = .006$). (3) Adding the flag to within-menu calibration (rule F vs. B) lowers the host's draw risk on the altered menu (the rate at which its samples miss the gold) from .652 to .610 at 50% coverage (95% interval excludes zero; Table 3).

that excluded candidate. Retaining the gold answer is the key design choice: the inserted candidate is never the gold answer, so every change in accuracy comes from the menu and can be traced, item by item, to a displaced answer. We obtain scores directly from four open-weight checkpoints across three families (0.6B to 4B parameters; Qwen2.5-1.5B-Instruct is our primary, or *host*, checkpoint) on controlled cloze menus and four established benchmarks.

Three findings emerge (Figure 1(c)). First, the menu effect can exceed the differences between models. On a single-token slice of CommonsenseQA, one swap lowers the host's accuracy from 66.0% to 21.6%, whereas the four checkpoints differ by at most 20.3 points on the same original menus, so a model comparison on a fixed menu partly measures the menu. A displacement identity accounts for the size of this effect across three checkpoints and four benchmarks, from 3.3 points on ARC-Challenge to 44.4 points on CommonsenseQA (correlation 0.998 over the 12 cells). Second, the ratio and a conventional *within-menu margin*, the gap between the two highest normalized listed scores, serve different targets: at the same retained count, the margin is the stronger selector of correct answers (81.1% vs. 68.4% correct on full-answer CommonsenseQA; likewise on ARC and OpenBookQA), whereas the ratio exposes menu sensitivity. Third, the ratio carries information about the altered menu: used as a flag in labelled calibration, it lowers the host's *draw risk* under that menu, the rate at which samples from it miss the gold answer, from .652 to .610 at 50% coverage (half of the items answered), and approximately halves it on the controlled cloze menus (from .209 to .104 on TinyStories).

Contributions.

- **A protocol that separates three targets.** A gold-free off-menu ratio and a gold-retaining one-distractor swap distinguish within-menu confidence, menu sensitivity, and answer correctness in one evaluation (Section 2). For single-token menus the ratio adds no model call, and exact-score accuracy is within 0.52 percentage points of plurality-vote accuracy over 12 sampled answers on controlled cloze menus (Section 2).
- **A measurement of menu sensitivity against model differences.** The swap lowers single-token CommonsenseQA accuracy by 41.8–46.8 points for all four checkpoints, about twice the largest original-menu gap, and ARC accuracy by 3.3–15.7 points. Because the inserted candidate is never the gold answer, the drop equals the displacement rate weighted by the accuracy of the displaced answers, which accounts for this variation (Section 4).
- **An equal-coverage test of correctness selection.** Items selected by the within-menu margin are 81.1% correct on full-answer CommonsenseQA, against 68.4% for *ratio-clean* items (no examined excluded candidate above the leader), with the same ordering on ARC and OpenBookQA. On CommonsenseQA, ratio-clean selection exceeds random subsets and the excluded-candidate-score control, the latter by 12.3 points; it also exceeds random subsets on RiddleSense (Section 5).
- **A calibration of draw risk on the altered menu.** Deflating the calibrated within-menu confidence of flagged items lowers the host's draw risk on the altered menu at 50% coverage from .652 to .610, with a 95% interval for the difference that excludes zero, and Qwen3-4B's by a smaller margin (.680 to .660; Section 6).

2 WHAT IS MEASURED?

Candidate scores and sampling. For prefix x, let $s(a \mid x)$ be a candidate score and $\mathcal{O}$ the presented menu. Define

$$q_{\mathcal{O}}(a) = \frac{\exp s(a \mid x)}{\sum_{b \in \mathcal{O}} \exp s(b \mid x)}, \quad \hat{a} = \arg\max_{a \in \mathcal{O}} s(a \mid x), \quad m = q_{\mathcal{O}}(\hat{a}) - q_{\mathcal{O}}(a_{(2)}), \tag{1}$$

where $a_{(2)}$ is the second-highest-scoring listed candidate, so m is the within-menu margin. For single-token answers, s is a token logit or log probability. The resulting q is the next-token distribution restricted to the menu. For multi-token answers, the score is *mean token log likelihood*, and its softmax is a distribution over these length-normalized scores. All candidate comparisons within an experiment use the same convention.

A count poll samples K times from $q_{\mathcal{O}}$ and chooses the most frequent answer by plurality vote (de Condorcet, 1785), with randomized ties. Its count margin estimates m, and its answer converges to $\hat{a}$ when the leader is unique. Exact argmax removes the Monte Carlo noise of these draws. In the single-token setting, $K = 12$ means one forward pass followed by 12 categorical draws; the costs of multi-token scoring and excluded-candidate construction are specified below. The two readouts agree closely. On the controlled cloze menus, exact argmax is 0.52 percentage points more accurate than the 12-draw plurality after integrating count ties (95% stem-cluster interval $[0.08, 0.96]$), and on single-token CommonsenseQA three checkpoints show approximate parity between the two (Appendix B.9); Appendices B.1 and B.4 analyze adaptive draw allocation and truncated polls.

An excluded-candidate comparison. Let $\mathcal{E}$ be a specified finite set of candidates absent from $\mathcal{O}$, and $w^* = \arg\max_{a \in \mathcal{E}} s(a \mid x)$. The diagnostic is the off-menu ratio τ, defined by

$$\log \tau = s(w^* \mid x) - s(\hat{a} \mid x). \tag{2}$$

Thus $\tau > 1$ means that an examined excluded candidate outranks every listed answer; we call such items *flagged* and items with $\tau \leq 1$ *ratio-clean*. A common normalization cancels from the ratio. The original-answer selection experiment retains $\tau \leq 1$ (ties are retained; there are none in its four evaluation sets). Some appendix analyses flag $\tau \geq 1$ instead, which differs only on exact ties (Appendix E). For multi-token scoring, the ratio compares exponentiated mean log likelihoods.

Cost. For single-token menus, the listed scores and the excluded search come from the same next-token forward pass: the excluded set is filtered from the top 4096 next-token candidates of that pass, so the ratio adds no model call. For the full-answer suites, listed options are scored with one prefix forward pass plus one batched forward pass over the option continuations, and the excluded candidates are model-native continuations of the same prefix (Appendix F.1).

The menu intervention. Replace the lowest-scoring nongold distractor by w^*, retaining the benchmark gold g; we call this intervention the *swap*. With $\mathcal{R}$ the resulting menu, measure the gold-mass change $\Delta = q_{\mathcal{O}}(g) - q_{\mathcal{R}}(g)$ and the change in argmax or count-poll accuracy. The gold label is used to construct and evaluate this controlled intervention; the raw diagnostic itself is gold-free. Because the menus are not part of the prompt context, both distributions are restrictions of the *same* per-item forward scores. Inserting a stronger distractor lowers the gold's normalized mass; deleting it raises that mass.

Three outcome variables. We report (i) original-menu answer accuracy, (ii) response displacement after the intervention, and (iii) draw risk on the altered menu (*re-menu draw risk*), the mean of $1 - \rho_i^{\mathcal{R}}$ among served items, where $\rho_i^{\mathcal{R}}$ is the empirical fraction of gold draws on the altered menu. The last quantity is a target distinct from original-answer error and from the error of the re-menu plurality answer.

3 Experimental design

Models, tasks, and candidate sets. The host checkpoint is Qwen2.5-1.5B-Instruct, with Qwen3-0.6B-Base, Qwen3-4B, and SmolLM2-1.7B-Instruct as additional checkpoints (Qwen et al., 2024; Yang et al., 2025; Allal et al., 2025). These are four checkpoints from three model families. Controlled cloze menus use TinyStories (Eldan & Li, 2023) and WikiText-103 (Merity et al., 2017): TinyStories v2 and v3 each have 1500 menu rows over the same 1498 distinct prefix–gold stems; WikiText has 1800 rows. Menus contain five single-token candidates chosen using corpus bigram counts. Excluded sets contain up to 16 filtered bigram continuations.

Established tasks are CommonsenseQA, ARC-Challenge, OpenBookQA, and RiddleSense (Talmor et al., 2019; Clark et al., 2018; Mihaylov et al., 2018; Lin et al., 2021). The single-token CommonsenseQA intervention uses 941 items (838 training and 103 validation rows); model-native excluded words are filtered from the top 4096 next-token candidates. This slice is distinct from the *full-answer* CommonsenseQA validation set of 1197 items used for ratio-based selection. The latter includes multi-token answers. ARC, OpenBookQA, and RiddleSense full-answer sets have 1460, 1000, and 1021 rows. Their scores use length normalization.

Selection and uncertainty. The original-answer comparison uses fixed evaluation halves for CommonsenseQA and RiddleSense (599 and 511 items) and official test slices for ARC and OpenBookQA (1165 and 500 items). Each comparator serves exactly as many items as the ratio-clean rule. The within-menu controls rank by m, the unnormalized top-two score gap, or the top softmax score; the excluded-candidate-score control ranks by the negative maximum excluded score. Equal-coverage cutoffs use evaluation *scores* only, without evaluation labels, in a fixed-pool comparison. We compute 5000 paired item bootstrap replicates conditional on the selected pools and checkpoint. Random-selection intervals are exact hypergeometric intervals for subsets of the fixed pool.

Labelled calibration of re-menu draw risk is a separate experiment: isotonic regression and flag-specific deflation are fit on one half and evaluated on the other. Repeated partitions measure split sensitivity. Cloze sampling comparisons integrate the count-tie randomization; pooled intervals resample 3298 prefix–gold clusters so that both TinyStories menu versions for a stem are resampled together. Full protocols and all measured conditions appear in the appendix.

4 How much can the menu change the result?

We first measure the magnitude of the swap effect. On the single-token CommonsenseQA slice, the host checkpoint's argmax accuracy falls from 66.0% to 21.6% when its strongest examined ex-

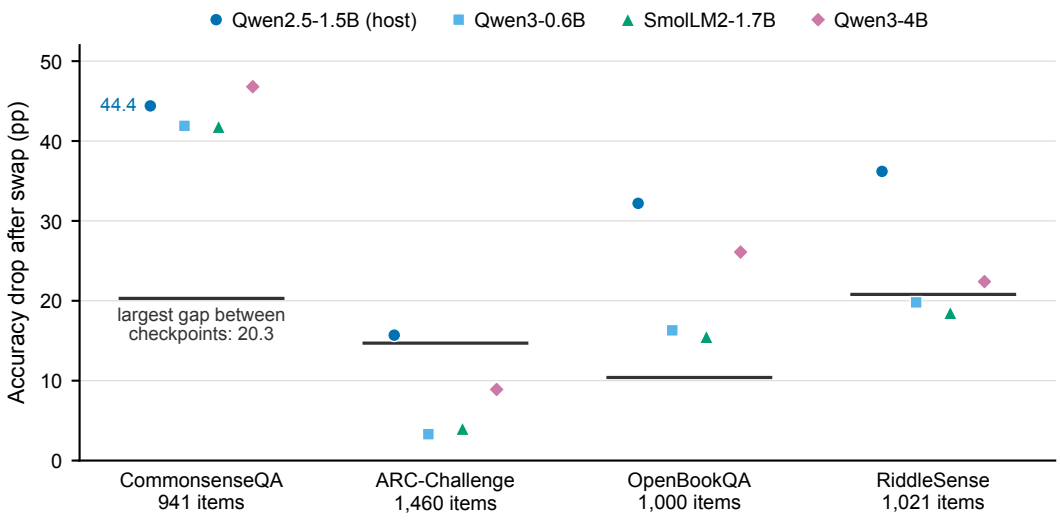

Figure 2: **Menu sensitivity can exceed model differences.** Points show each checkpoint's decrease in argmax accuracy after its own strongest excluded candidate replaces a distractor. Horizontal bars show the largest between-checkpoint gap on the identical original menus. The 941-item single-token CommonsenseQA slice is separate from the full-answer selection experiment. Item counts are shown under each suite; the host's CommonsenseQA drop (44.4 pp) and the largest CommonsenseQA gap between checkpoints (20.3 pp) are labelled directly. Item-bootstrap intervals are in Appendix F.5.

cluded candidate replaces a distractor.[1] The 44.4-point change exceeds the largest original-menu gap between the four checkpoints, 20.3 points. The other checkpoints lose 41.8–46.8 points. Figure 2 shows the full comparison across suites: on ARC the host loses 15.7 points and the other three checkpoints lose 3.3–8.9 points. The effect size thus varies with task and checkpoint, and the identity below accounts for this variation.

Why the drop is large: a displacement identity. The size of the drop follows from two structural facts of the intervention (Appendix A.2). First, removing the lowest-scoring nongold option never removes the listed leader, so an item's answer changes only if w^* outscores it. Second, w^* is never the gold answer, because it is drawn from candidates absent from a menu that already contains the gold. With binary indicators o_i for original-menu correctness, d_i for displacement of the leader by w^*, and r_i for altered-menu correctness, every item satisfies $r_i = o_i(1-d_i)$, so the accuracy change is $\Pr(d)\,\mathbb{E}[o \mid d = 1]$: the displacement rate weighted by the accuracy of the displaced answers. Both facts hold on every item of the 12 cells (three checkpoints $\times$ four suites).

Across these cells, displacement rate times overall accuracy tracks the measured shift with correlation 0.998 and mean absolute error 0.85 points (Table 1; Appendix F.5). Qwen3-4B, the replication checkpoint in Figure 2, is analyzed separately (Appendix F.8). The displacement rate is high on CommonsenseQA (.72 for the host) and much lower on ARC (.33), which explains the task dependence. Because displacement rates differ, aggregate altered-menu accuracies can change order even though no item's correct/incorrect relation between two checkpoints reverses: on ARC, the host's 11.7-point larger shift outweighs its 4.7-point lead over SmolLM2 (Appendix A.2). On the 1481 TinyStories v3 items whose w^* shares no first token with a listed answer, the mean gold mass predicted after the intervention is .619 and the realized 12-draw mean is .621, with item correlation .94 (same appendix).

When does the menu effect exceed model differences? The identity also gives a simple condition. With shift $\approx P(d) \cdot \mathrm{Acc}$, a checkpoint's shift exceeds the largest original-menu gap G between the checkpoints when its displacement rate exceeds G/Acc. For the three checkpoints of Table 1, this condition agrees in all 12 cells with a joint item bootstrap that resamples the same items for

[1]This result uses the draw-free argmax readout on all 941 items (Appendix F.5). The accuracy decreases from 12-draw count polls on the same slice are 43.4–44.2 points for the host (Appendix A.2 and Table 17).

Table 1: **The displacement identity across the 12 checkpoint–task cells.** Acc is original-menu argmax accuracy, $P(d)$ the rate at which w^* displaces the listed leader, Pred. the product $P(d) \cdot \mathrm{Acc}$, and Shift the measured accuracy decrease after the intervention; Pred. empirically tracks Shift across cells. Within a suite all checkpoints score identical menus of the listed size; CommonsenseQA is the single-token slice. Bootstrap intervals in Appendix F.5.

		Qwen2.5-1.5B (host)				Qwen3-0.6B				SmolLM2-1.7B			
Suite	Items	Acc	$P(d)$	Pred.	Shift	Acc	$P(d)$	Pred.	Shift	Acc	$P(d)$	Pred.	Shift
		%		pp	pp	%		pp	pp	%		pp	pp
CommonsenseQA	941	66.0	.721	47.5	44.4	45.7	.947	43.3	41.9	48.0	.877	42.1	41.8
ARC-Challenge	1,460	46.9	.333	15.6	15.7	35.9	.103	3.7	3.3	42.2	.106	4.5	4.0
OpenBookQA	1,000	42.4	.755	32.0	32.2	33.7	.468	15.8	16.3	36.0	.436	15.7	15.5
RiddleSense	1,021	48.3	.820	39.6	36.2	27.5	.717	19.7	19.8	33.1	.556	18.4	18.5

every checkpoint: the condition holds, and the 95% interval of shift minus gap lies above zero, for every checkpoint on CommonsenseQA and OpenBookQA and for the host alone on ARC and RiddleSense. For the host on CommonsenseQA, the shift exceeds the gap by 24.1 points (95% interval $[20.3, 27.7]$). The shift is also stable under a change of prompt format: with five format-only demonstrations, the host's shift on a 95-item CommonsenseQA subset is 35.8 points, the same as with the original prompt (Appendix F.9).

Are the displacing candidates plausible answers? Exact match with the gold is zero by construction, but w^* might still be a legitimate answer that the menu omits. A forced pairwise judge (Llama-3.1-8B-Instruct) chooses which of two candidates better answers the question (Appendix G.1). On probe pairs of a gold answer and a listed distractor, its direction accuracy (the rate at which it prefers the gold, averaged over both presentation orders) is 76–83%. On flagged items, it rates w^* at least as good as the gold on 13.4% (ARC), 19.5% (CommonsenseQA), 20.9% (RiddleSense), and 32.7% (OpenBookQA) of items. The judge thus identifies a minority of omitted candidates as plausible alternatives and ranks most strictly below the gold.

5 Which score selects correct answers?

We next ask which gold-free score best selects correct original answers for abstention. Figure 3 compares the ratio-clean pool with two matched-size alternatives, and Table 2 lists all five selectors. On full-answer CommonsenseQA, the ratio retains 212 of 599 evaluation items and achieves 68.4% accuracy. Selecting the same number by the within-menu margin achieves 81.1%, a ratio-minus-margin difference of -12.7 points (95% interval $[-19.0, -6.5]$), so the margin is the stronger correctness selector. The ratio exceeds the excluded-candidate-score comparator by 12.3 points. That control ignores the listed scores, so the within-menu margin is the relevant baseline for abstention.

The within-menu margin also selects more accurate answers on ARC and OpenBookQA, with ratio-minus-margin differences of -4.3 points $[-7.1, -1.4]$ and -15.4 points $[-27.2, -3.9]$; on RiddleSense the ratio-minus-margin difference is -1.0 points $[-12.4, 10.4]$. Top-score confidence and the unnormalized score gap give the same pattern (Table 2).

Against random subsets of equal size, ratio-clean selection raises accuracy on CommonsenseQA and RiddleSense: 68.4% and 68.0% versus fixed-pool random expectations of 61.4% and 50.7% (upper-tail $p = .006$ and $p < .001$). The unit threshold retains 35.4%, 68.2%, 23.4%, and 19.6% of the four pools in column order (Appendix F.6). The results thus separate the diagnostic targets: the within-menu margin selects correct original answers, and the off-menu ratio measures menu sensitivity.

6 Calibrating draw risk on the altered menu

A different task is to estimate draw risk after the specified menu intervention. Let $\hat{\rho}(m)$ be an isotonic fit of within-menu gold-draw frequency. The within-menu rule B ranks by $\hat{\rho}(m)$. The

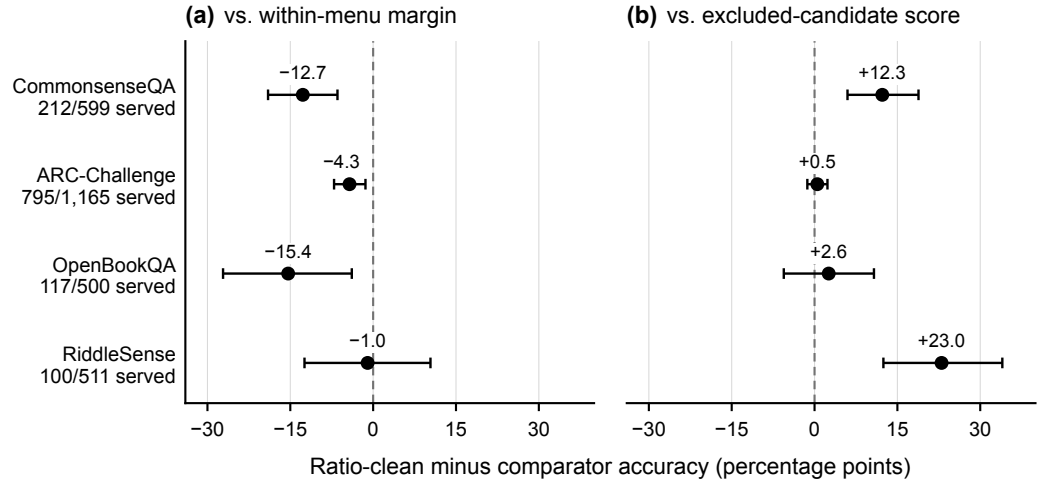

Figure 3: **Equal-coverage selection against two comparators.** Original-answer accuracy of ratio-clean items minus a comparator at exactly equal served count. Positive values favor the ratio. The within-menu margin is stronger on CommonsenseQA, ARC, and OpenBookQA; the ratio's point estimate exceeds the excluded-candidate score in all four suites, with intervals above zero on CommonsenseQA and RiddleSense. Error bars are 95% paired item bootstrap intervals, conditional on the fixed selected pools and host checkpoint. Numbers above markers give the point estimates in percentage points; row labels give served/evaluation counts. All four suites use full-answer length-normalized scores; this CommonsenseQA pool is separate from the 941-item intervention slice.

Table 2: **Original-answer accuracy (%) of five selectors at identical served count within each column** (host checkpoint, full-answer length-normalized scores). The ratio-clean rule keeps items with $\tau \leq 1$. Each selector below it ranks items by a within-menu score, or by the maximum excluded score (lowest first), and keeps as many items as the ratio-clean rule. Bold marks the highest selector accuracy in each column (ties are all bold). The random rows give the exact hypergeometric mean and 95% range for random subsets of the same size from the fixed evaluation pool, and the upper-tail probability p of the ratio-clean accuracy. Full protocol in Appendix F.10.

Selector	CommonsenseQA	ARC-Challenge	OpenBookQA	RiddleSense
Served / evaluation items	212/599	795/1,165	117/500	100/511
Ratio-clean ($\tau \leq 1$)	68.4	46.5	45.3	68.0
Within-menu margin m	**81.1**	50.8	60.7	69.0
Unnormalized score gap	**81.1**	**50.9**	**62.4**	**71.0**
Top-score confidence	**81.1**	49.8	60.7	70.0
Excluded-candidate score	56.1	46.0	42.7	45.0
Random subset of equal size, mean	61.4	46.5	45.8	50.7
95% range	[56.1, 66.5]	[44.5, 48.4]	[37.6, 53.8]	[42.0, 59.0]
upper-tail p of ratio-clean	.006	.518	.590	$<$.001

flag-deflated rule F ranks by

$$\hat{\rho}_{\mathcal{R}} = \hat{\rho}(m)\, \eta(\text{flag}, \text{margin bin}), \tag{3}$$

where η is the median on the fit half of the re-menu/within-menu gold-frequency ratio. The flag uses no gold label at decision time; fitting η and $\hat{\rho}$ uses the labels of the fit half.

Figure 4 shows both rules across coverage on single-token CommonsenseQA. At 50% coverage, the host's risk decreases from .652 to .610, and Qwen3-4B's from .680 to .660.

Table 3 ties the reduction to the unflagged pool. The host's reduction has a 95% interval that excludes zero: with a flag rate of .721 over all items, the unflagged evaluation-half pool can supply up to .284 coverage, and deflation routes service through that pool (Appendix E.8). The flag is also informative

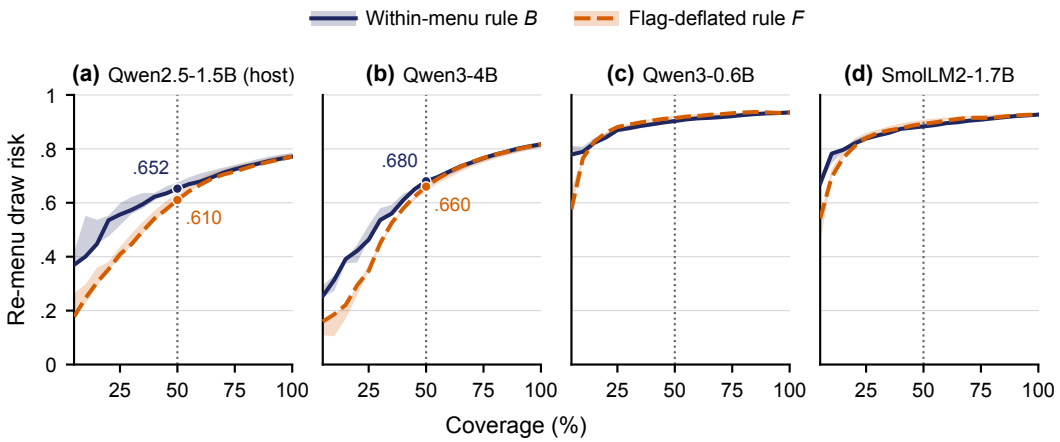

Figure 4: **Re-menu draw risk of the within-menu and flag-deflated rules.** Lower is better for empirical gold-draw error after the menu intervention. Lines show the primary fit/evaluation partition; shading spans the minimum and maximum of three partitions. All 20 measured coverages are shown. The dotted line marks 50% coverage, and dots mark the 50%-coverage values cited in the text. Evaluation sizes are 471, 470, 471, and 410 in panels (a)–(d); the SmolLM2 scoring basis contains 820 items, and one numerically degenerate Qwen3-4B item is excluded from both arms.

Table 3: **Re-menu draw risk at 50% coverage on single-token CommonsenseQA** (evaluation half; lower is better). B is the within-menu rule and F the flag-deflated rule. $F - B$ is the mean paired bootstrap difference with its 95% interval (2000 resamples, conditional on the fit half); bold marks the difference whose interval excludes zero. The flag rate is computed over all items (Table 17); the clean cap is the largest coverage the unflagged pool can supply, one minus the evaluation-half flag rate. Full protocol in Appendix E.8.

Checkpoint	Flag rate (all items)	Risk B	Risk F	$F - B$ [95% interval]	Clean cap
Qwen2.5-1.5B (host)	.721	.652	.610	**−.042** [−.067, −.018]	.284
Qwen3-4B	.773	.680	.660	−.016 [−.037, +.003]	.238
Qwen3-0.6B	.947	.903	.914	+.011 [−.002, +.025]	.055
SmolLM2-1.7B	.946	.883	.893	+.011 [−.004, +.024]	.054

about the altered menu itself: for the host, w^* becomes the pooled 12-draw answer on .944 of flagged items, and .967 of the items it takes over are flagged (Table 17). Qwen3-4B, with a clean cap of .238, shows a smaller mean reduction; Appendix I discusses the remaining two rows. The controlled cloze tasks show larger reductions in re-menu draw risk: at 50% coverage, flag deflation lowers it from .209 to .104 on TinyStories v3 and from .227 to .114 on WikiText-103 (Appendix E.1, Table 15). The selection result of Section 5 and this calibration result optimize different outcomes: the within-menu margin selects correct original answers, and the flag adds information about risk under the altered menu.

7 Related Work

What multiple-choice scores measure. Likelihood scoring is standard in language-model evaluation (Brown et al., 2020; Hendrycks et al., 2021; Robinson & Wingate, 2023), but it is sensitive to the scoring and presentation protocol. Contextual calibration addresses label priors (Zhao et al., 2021); option-order and identifier studies show additional biases (Pezeshkpour & Hruschka, 2024; Zheng et al., 2024). Mechanistic work distinguishes answering from assembling an option identifier (Wiegreffe et al., 2025). A closely related study decomposes choice-driven and question-driven scores across multiple prompting formats (Cho et al., 2026). Our fixed-prefix candidate replacement probes a different axis: a candidate excluded from the scored menu, without changing the prompt's option order. Generative substitutes for forced choice, such as answer matching and judge-

based scoring (Es et al., 2024; Zheng et al., 2023), replace the protocol rather than isolate what it hides; our measurements keep the forced-choice record intact. Studies of popular options and absent correct answers further motivate reporting menu construction (Abdallah et al., 2026; Mensa et al., 2026); here the gold answer stays on the menu throughout.

Confidence, sampling, and selection. Self-consistency aggregates open-ended reasoning paths (Wang et al., 2023); confidence-guided stopping and reasoning-aware sampling reduce generation budgets (Aghazadeh et al., 2025; Wan et al., 2024). Our fixed-menu categorical draws use directly accessible scores, so one forward pass supports all draws. Model self-evaluation, verbalized confidence, semantic uncertainty, and process verification target other aspects of reliability (Kadavath et al., 2022; Lin et al., 2022; Kuhn et al., 2023; Lightman et al., 2023). Standard calibration and selective-prediction methods supply the relevant correctness controls (Guo et al., 2017; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; 2019). Conformal and risk-controlling methods formalize guarantees under specified assumptions (Angelopoulos & Bates, 2021; Bates et al., 2021; Angelopoulos et al., 2021; Quach et al., 2024).

8 Conclusion

Within-menu confidence and menu sensitivity are distinct, separately measurable quantities. A gold-free off-menu ratio and a gold-retaining one-distractor swap expose menu sensitivity that can exceed model differences (44.4 vs. 20.3 points on single-token CommonsenseQA), and a displacement identity accounts for its size across tasks. The within-menu margin is the stronger selector of correct original answers, and the off-menu flag lowers calibrated risk under the altered menu for the host checkpoint. Evaluations should report the menu, the excluded search, and the calibration target alongside accuracy; compare model differences with the size of a gold-retaining menu change; and use the within-menu margin for abstention and the off-menu ratio for menu completeness.

8.1 Limitations and future work

The altered menus are checkpoint-specific, and prediction–draw agreement checks score normalization and sampling. Effect direction follows the intervention by construction; the empirical comparison is its magnitude relative to model differences. Results condition on fixed datasets and checkpoints, and calibration is empirical for the specified intervention. Appendix I details these limits. We identify four directions for future work. *Scale and access:* extending the protocol from small open-weight checkpoints with score access to larger, API-only, and independently retrained models, on benchmarks free of contamination concerns. *Candidate search and scoring:* replacing the finite, filtered excluded set and length-normalized scores with broader generative search and complete-answer probabilities, so that normalization, answer length, and tokenization can be isolated. *Semantic validity:* complementing the single model judge with human annotation, to measure how often an omitted candidate is a legitimate answer. *Selection and calibration:* combining the within-menu margin with the off-menu flag, and adding distribution-free risk control to a calibration benefit that currently varies across checkpoints.

AI Use Statement

During the project, vibe coding tools (Codex and Claude Code) were used to help develop the code for the methods and experiments. All such work was conducted under human supervision and underwent human review. During manuscript preparation, AI tools were used to polish and revise the text and figures.

Ethics Statement

This work evaluates language models using publicly available corpora, benchmarks, and model checkpoints. It involves no human-subject experiments. Ethical considerations include those associated with the source data and models and with the interpretation of benchmark scores.

Reproducibility Statement

The supplementary material (a zip archive) provides code, configurations, launch and analysis scripts, and recorded results. Its README maps tables and figures to their source files and analysis scripts. Experimental settings are specified in the main text and the appendix.

References

Abdelrahman Abdallah, Mohammed Ali, Bhawna Piryani, Mahmoud Abdalla, and Adam Jatowt. Large language models systematically favor popular options: Evidence and mitigation across MCQs. *arXiv preprint arXiv:2608.29257*, 2026.

Ehsan Aghazadeh, Ahmad Ghasemi, Hedyeh Beyhaghi, and Hossein Pishro-Nik. CGES: Confidence-guided early stopping for efficient and accurate self-consistency. *arXiv preprint arXiv:2511.02603*, 2025.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, et al. SmolLM2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025. URL https://arxiv.org/abs/2502.02737.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candes, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. arXiv:2005.14165.

Gyeongje Cho, Yeonkyoung So, and Jaejin Lee. Choices speak louder than questions. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7ae2b0eedc560e3afbffd68445b8a220-Abstract-Conference.html.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Nicolas de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. De l'imprimerie royale, 1785.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Ronen Eldan and Yuanzhi Li. TinyStories: How small can language models be and still speak coherent English? *arXiv preprint arXiv:2305.07759*, 2023. URL https://arxiv.org/abs/2305.07759.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the ACL (System Demonstrations)*, pp. 150–158, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. *International Conference on Machine Learning*, 2019. URL https://arxiv.org/abs/1901.09192.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *International Conference on Machine Learning (ICML)*, 2017. arXiv:1706.04599.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2009.03300.

Saurav Kadavath, Tom Conerly, Amanda Askell, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *International Conference on Learning Representations (ICLR)*, 2023. Spotlight; arXiv:2302.09664.

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

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. RiddleSense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1504–1515, 2021.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022. URL https://arxiv.org/abs/2205.14334.

Enrico Mensa, Lorenzo Zane, Calogero Jerik Scozzaro, Matteo Delsanto, Tommaso Milani, and Daniele Paolo Radicioni. Easy to complete, hard to choose: Investigating LLM performance on the ProverbIT benchmark. *arXiv preprint arXiv:2608.04670*, 2026.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1609.07843.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2381–2391, 2018.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024. arXiv:2308.11483.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.10193.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Joshua Robinson and David Wingate. Leveraging large language models for multiple choice question answering. In *International Conference on Learning Representations (ICLR)*, 2023.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 4149–4158, 2019. arXiv:1811.00937.

Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. Reasoning aware self-consistency: Leveraging reasoning paths for efficient LLM sampling. *arXiv preprint arXiv:2408.17017*, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2203.11171.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hanna Hajishirzi, and Ashish Sabharwal. Answer, assemble, ace: Understanding how LMs answer multiple choice questions. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c248154176c08147e82c0b30961604f7-Abstract-Conference.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. *International Conference on Machine Learning*, 2021. URL https://arxiv.org/abs/2102.09690.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.03882.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks*, 2023.

A GUIDE TO THE SUPPLEMENTARY EVIDENCE

This appendix separates exact relationships from empirical performance. The first two subsections derive finite-sample poll probabilities and the conditional re-menu collapse identity. The remaining sections organize all recorded positive and negative results by research question. File names cited beside individual results (`results/...`, `scripts/...`, `data/...`) refer to files in the supplementary material; Appendix H describes its layout.

Question	Section (page)
What follows from the score and sampling definitions?	A (13)
When do draws differ from exact-score decisions?	B (16)
How sensitive is the flag to the candidate menu?	C (21)
Which omissions and errors can the diagnostic miss?	D (26)
When does the flag help selective serving?	E (38)
Which effects transfer across established benchmarks?	F (45)
Does an excluded candidate have semantic validity?	G (52)
Which inputs and implementation conventions are used?	H (53)
What are the scope conditions and limitations?	I (54)

Notation and checkpoint names. The appendix uses short codes: g1 (or g_1) is the host checkpoint Qwen2.5-1.5B-Instruct, g2 (or g_2) is Qwen3-0.6B-Base, smol is SmolLM2-1.7B-Instruct, and q4b is Qwen3-4B; "family" refers to one of these checkpoints, and a *leg* (or case) is a checkpoint–suite combination. The controlled cloze suites are ts v2 and ts v3 (TinyStories menu versions) and wt v2 (WikiText-103); CSQA is CommonsenseQA, and CSQA-941 is the single-token intervention slice. R0 denotes the gold-retaining intervention of Section 2, which replaces the lowest-scoring distractor by w^*. The "threat" is the off-menu ratio τ; "under-inclusion" refers to items on which an excluded candidate outscores the listed leader, and its "depth" is the displacement rate $P(w^* \mid \text{re-menu answer})$. A "hostage" fraction is the share of a checkpoint's original-menu accuracy that the intervention removes, $\mathrm{shift}/\mathrm{acc}$; it describes sensitivity to this intervention, not answer quality, and items whose served answer the intervention displaces are called hostage items. Serving "cov$c$" denotes coverage $c\%$. Other abbreviations: LL, length-normalized log likelihood; MC, multiple choice; OOS, out-of-sample; ECE, expected calibration error; AUC, area under the ROC curve; PAVA, the pool-adjacent-violators algorithm for isotonic regression; MAE, mean absolute error; BIC, Bayesian information criterion; pp, percentage points. The exclusion $P(w^*{=}\text{gold}) = 0$ holds by construction in every benchmark cell, because w^* is drawn from candidates absent from a menu that already contains the gold; item-level checks confirm it.

Common interpretation. The threat ratio compares the strongest examined excluded candidate with the presented leader on a common score scale: with common-normalizer probabilities, $\tau = p^{(S)}_{\text{off}}/p^{(S)}_{\text{lead}}$. A unit threshold signals an omission relative to that candidate set; it does not certify a correct alternative answer. The re-menu intervention retains the gold label and replaces a distractor. Selective-serving results are empirical. Re-menu calibration separates labelled fit and evaluation samples; equal-coverage original-answer comparisons choose their cutoffs from evaluation scores without labels. Across suites, comparisons retain different tokenization, candidate-construction, and generator conditions; their effect sizes are not pooled as independent model replications.

A.1 EXACT FINITE-SAMPLE POLL PROBABILITIES

For fixed menu probabilities $p_1, \ldots, p_M$, let $T(n) = \{j : n_j = \max_l n_l\}$ and let j^* be the unique most probable option. With uniform tie breaking, the exact probability that the count poll agrees with j^* is

$$\sum_{\substack{n_1+\cdots+n_M=K \\ n_j \geq 0}} \frac{K!}{\prod_j n_j!} \prod_j p_j^{n_j} \frac{\mathbf{1}\{j^* \in T(n)\}}{|T(n)|}.$$

Replacing j^* in the final factor with the benchmark gold gives expected poll accuracy. Agreement with the model argmax and agreement with gold are different quantities; the latter has no general dominance ordering at finite K.

The count-poll deficit at $K{=}12$ is a finite-sample consequence of the mass-margin identity, and it is *predictable item-by-item* from the per-item multinomial (enumerate all $\binom{16}{4}{=}1820$ count vectors over $K{=}12$ draws, integrating over uniform tie-breaking; `results/asm_k_lag_predict.json`): the predicted P(count-argmax $\neq$ logit-argmax) is $.024/.030/.018$ versus its realized value $.024/.035/.012$ (ts v2 / ts v3 / wt v2; per-suite deviation $\leq$ .7pp), and the predicted expected one-forward gap $E[\ell{-}c]$ is $+.66/+.91/+.48$pp vs. realized seeded $+.27/+1.07/+.28$pp (pooled prediction $+.67$pp vs. realized $+.52$pp, both under the shared-stream protocol; per-suite deviations of at most .39pp / .16pp / .20pp — the pooled comparison is the realization-stable one). The small exact-score advantage over the 12-draw poll reported in Section 2 is therefore predicted by the finite-K multinomial at $K{=}12$, a value that can be computed from the model's own menu masses before sampling. The realized seeded disagreement rates of Table 9 are $.024/.033/.013$ (ts v2/ts v3/wt v2); integrating out ties instead gives $.0240/.0322/.0139$. These convention-specific realizations share the same exact multinomial prediction. Against a single draw, the accuracy gains are $+5.33/+5.47/+2.61$ percentage points (ts v2/ts v3/WikiText v2).

For these three controlled cloze suites, the single-draw accuracy advantages are $+5.5/+5.3/+2.6$ percentage points (ts v3/ts v2/wt v2), with intervals $[4.0, 6.9]$, $[3.9, 6.6]$, and $[1.6, 3.7]$. The advantage over the 12-draw poll is smaller and depends on menu construction: the stem-cluster two-sided p value is .026 for the ts v3 plus WikiText combination and .146 for ts v2 plus WikiText. These empirical comparisons do not follow from the agreement identity alone. Established-menu cases include parity and a SmolLM2 reversal (Appendix B.9).

A.2 Why the aggregate re-menu ordering changes while the item ordering does not

The ARC coexistence – the observed ranking g1 $>$ SmolLM2 $>$ g2 flipping to SmolLM2 $>$ g2 $>$ g1 in *aggregate* re-menued accuracy while no item's gold-relation ever reverses – has two different explanations. The absence of strict item-level reversals follows from the structure of the R0 intervention, whereas the aggregate change depends on the measured displacement rates (`results/asm_stakes_decomp.json`). For each family f and item i, write $o_{f,i} \in \{0,1\}$ for observed-menu answer correctness, $d_{f,i} \in \{0,1\}$ for "w_f^* displaces the in-menu answer on the R0 re-menu," and $r_{f,i} \in \{0,1\}$ for re-menu correctness. Two structural facts hold exactly (verified item-by-item, 0 violations in *each* of the twelve family$\times$suite cells):

- (P1) *the R0 re-menu preserves the answer unless displaced*: the drop-weakest rule removes the lowest-LL non-gold option, so the in-menu LL-argmax is unchanged, and $r_{f,i} = o_{f,i}$ whenever $d_{f,i} = 0$;
- (P2) *the displacing candidate is never the gold*: w_f^* is drawn from candidates absent from a menu that already contains the gold, so by construction the re-menu answer on displaced items is incorrect (an item-level check confirms frequency 0.000 in all twelve cells), and $r_{f,i} = 0$ whenever $d_{f,i} = 1$.

Together these give the *collapse identity* $r_{f,i} = o_{f,i}(1 - d_{f,i})$: the re-menu operator can turn a correct answer into an incorrect one, and can do nothing else.

Why strict item-level reversals are impossible under (P1)–(P2). With binary correctness, $r_f \in \{o_f, 0\}$ pointwise. For any pair $(\mathrm{rel}_{\mathrm{obs}}(i) = \mathrm{sign}(o_a - o_b) = +1)$ we must have $o_a{=}1, o_b{=}0$, hence $r_a \in \{0,1\}$ and $r_b{=}0$, so $\mathrm{rel}_{\mathrm{re}}(i) \in \{0,1\}$ – never -1. Symmetrically $\mathrm{rel}_{\mathrm{obs}} = -1 \Rightarrow \mathrm{rel}_{\mathrm{re}} \in \{-1, 0\}$. A strict pair-relation reversal ($\mathrm{rel}_{\mathrm{obs}} \cdot \mathrm{rel}_{\mathrm{re}} < 0$) is therefore *impossible* for any displacement pattern; the only per-item movements are "collapse" events ($o{=}1, d{=}1$, the item's correct answer is displaced) and "zeroed-pair" events ($o_a{=}o_b{=}1$, both displaced). These events are abundant – on ARC g1 has 229 collapse events (15.7% of items, exactly the shift in points), g2 48 and SmolLM2 58; for the g1-vs-SmolLM2 pair there are 94 collapses of g1 while SmolLM2 was wrong, 25 the other way, and 22 both-right both-displaced items – so "0 reversals" is not vacuous; it holds because a collapse is a zero, never a substitution. The same arithmetic explains *why* the ARC aggregate order can flip: the re-menu lets $g1$'s deeper under-inclusion ($P(d){=}.333$ vs. $.106/.103$), which sets the shift (shift$=P(d)\,\mathbb{E}[o \mid d]$, so $15.7/4.0/3.3$pp), outweigh its 4.7pp accuracy advantage.

Why aggregate rankings can change. Summing the identity, $r\text{-acc}_f = o\text{-acc}_f - \text{shift}_f$ with $\text{shift}_f = P(d_f)\,\mathbb{E}[o_f \mid d_f{=}1]$, so for any pair

$$\Delta_r(a,b) \;=\; \Delta_o(a,b) \;-\; \big(\text{shift}_a - \text{shift}_b\big).$$

The aggregate re-ranking of (a,b) flips if and only if $|\text{shift}_a - \text{shift}_b| > |\Delta_o(a,b)|$ and the shift difference has the same sign as $\Delta_o(a,b)$. Equivalently, $\Delta_o(a,b)\Delta_r(a,b) < 0$. Numerically (pp): ARC g1-vs-SmolLM2 $\Delta_o{=}+4.7$, shift-gap $15.7{-}4.0{=}11.7$, so $\Delta_r{=}-7.0$ – the 11.7pp menu-completeness penalty gap exceeds the 4.7pp raw accuracy gap, and the pairwise ordering after this menu intervention reverses; g1-vs-g2: $+11.0 - (15.7{-}3.3){=}-1.4$ (flip); SmolLM2-vs-g2: $+6.3 - (4.0{-}3.3){=}+5.6$ (no flip, the observed order survives). On CSQA every shift gap is tiny ($\leq$ 2.6pp) relative to every observed gap ($\geq$ 2.3pp); these two bounds alone do not determine a pair's outcome, but the recorded pair-specific contrasts do not reverse. The aggregate order change on ARC, and its absence on CSQA follow from the same formula with the two suites' displacement rates substituted. In the near-orthogonal regime where displacement does not select for difficulty ($\mathbb{E}[o_f \mid d_f{=}1] \approx o\text{-acc}_f$; ARC, mean $|\text{err}| = .006$), the formula collapses to the closed form $r\text{-acc}_f \approx o\text{-acc}_f\,(1 - P(d_f))$, so the aggregate re-order is the order of the *surviving* accuracy: g1 $.469(1 - .333) = .313$ vs. SmolLM2 $.422(1 - .106){=}.377$ — the family with the deepest under-inclusion is discounted by its own displacement rate. On CSQA $\mathbb{E}[o_f \mid d_f]{<}o\text{-acc}_f$ (the displaced items had weaker leaders), so the simple form overstates the drop and only the exact identity is quoted there — a stated applicability boundary of the closed form. Equivalently, the fraction of a family's served accuracy that this intervention removes (the "hostage" fraction) is $\text{shift}_f / o\text{-acc}_f = P(d_f)\,\mathbb{E}[o_f \mid d_f] / o\text{-acc}_f$: $\approx .33/.09/.09$ on ARC (g1/SmolLM2/g2), $.76/.43/.48$ on OpenBookQA and $.67/.87/.92$ on CSQA — on the deep-under-inclusion suites the intervention removes 87–92% (CSQA) and 43–48% (OpenBookQA) of the two *weaker* families' original-menu accuracy, so on these menus a cross-model comparison is strongly affected by menu completeness as well as by model quality.

Sampling interpretation. Treating N items as independent draws, observing 0 reversals gives the following binomial bound: the per-item strict-reversal rate is bounded by $1 - .05^{1/N}$ at 95%: $\leq .21\%$ (ARC, $N{=}1460$) and $\leq .32\%$ (CSQA, $N{=}941$) — the theorem predicts exactly 0, and the data place the empirical rate at or below a few tenths of a percent per item.

Finite-sample count readouts. The identity is exact for the draw-free LL read; under the sampled $K{=}12$ count read (g1/CSQA, the only real-draw read of this re-menu decomposition; the all-family CSQA count read is Appendix B.9) the aggregate decomposition still holds *exactly* (shift $= P(d)\,\mathbb{E}[o \mid d] = 44.2$pp) while the per-item identity is violated on at most $2/941$ items (0.2%, the draw-noise floor) — switching readouts changes item-level invariance from exact to approximate within sampling noise, and no further.

Score-level check of the intervention. The same intervention also has a transparent score-level explanation (referenced in Section 4). On the 1481 TinyStories v3 items whose w^* shares no first token with a listed answer (the other 19 of the 1500 items are excluded), the predicted mean gold mass after the intervention is .619 and the realized 12-draw mean is .621. Their item correlation is .94. These values agree because both use the same underlying scores; they establish that the measured response follows the defined menu change.

When the conditions fail. The item invariance requires exactly (P1) and (P2). (P1) is a property of the R0 construction (keeping the in-menu argmax); a re-menu rule that removed the argmax ("drop the strongest," or random removal) would break it. (P2) holds by construction in the benchmark protocol (w^* is absent from a menu that contains the gold), and an item-level check confirms $P(w^*{=}\text{gold}) = .000$; if serving the displaced candidate were correct with probability $q{>}0$, strict reversals would appear through the both-displaced and single-displaced overlap: at $q{=}.25$ a deterministic seed gives a mean of $40.5/38.0/47.1$ reversals per 200 draws on CSQA (both-displaced overlap 623–796 items) and $6.2/7.4/5.4$ on ARC (overlap 96–112) — the coexistence is fragile to w^* ever being gold and would then signal it. Since w^* cannot be the item gold under this construction, the relevant case has $q{=}0$ and the invariance is exact, so "the menu-completeness stimulus can move *aggregate* model scores without reordering which family is correct on any item" is the precise statement under conditions (P1)–(P2).

B SAMPLING AND CONFIDENCE WITHIN A FIXED MENU

B.1 ON-DEMAND REALLOCATION YIELDS NO MEASURED GAIN

An adaptive early-exit scheduler (draw 4; answer if $\delta(4) \geq \tau_a$; top-up hostage/ambiguous items to 8/12) was implemented and measured on the real paths at matched total budget. The accuracy-vs-budget frontier of the *uniform* schedules is strictly monotone (k=2 : .924, 4 : .965, 6 : .977, 8 : .980, 12 : .985 at c=.5, ts v3), and every adaptive configuration lies *on* the frontier at its own budget, never above it: on this benchmark the uniform batch schedule is budget-optimal. The reason is mechanism, not engineering: extra draws sharpen both the margin estimate and the pooled answer symmetrically, so reallocation provides no accuracy gain at fixed budget. (Adaptive and early-stopped draw scheduling has been proposed to save draw budget on self-consistency pipelines (Aghazadeh et al., 2025; Wan et al., 2024); on this generator and benchmark family we measure no accuracy gain of adaptive reallocation beyond the uniform schedule's budget frontier.)

B.2 SUITE DEPENDENCE OF THE POOLED ACCURACY COMPARISON

The discordant-pair tables from the per-suite-reseed realization (`results/asm_answer_gap_meta_exact_v2.json`) combine to OR 1.61 (Woolf z=2.28, p=.022; I^2=0%), consistent with the pooled 60/37 binomial (p=.025) — but these are the arithmetic of one random-number protocol under independent-row assumptions; the stem-cluster interval of Section 2 does not depend on a single realization and is the one we report. The serving and calibration analysis (Section B.7) is independent of this answer-level cell.

B.3 TEMPERATURE SENSITIVITY OF THE DOMINANCE CLAIM

The masked read is drawn at temperature T=1; in self-consistency pipelines higher T is common, and the exact argmax is temperature-invariant while the count poll's noise grows with T (softmax/T flattens). To make the dominance claim robust to that choice we re-ran the identical ts v3 K=12 draws at T=2 (`results/asm_temp2_consume.json`): the one-forward accuracy is unchanged by construction (.927), while the count poll drops from .915 to .894 (−2.1pp) and a *single* draw from .872 to .779 (−9.3pp) — the draw-free advantage over the poll *widens* from +1.2pp to +3.3pp, entirely from the sampling side. The higher-temperature operating point only sharpens the dominance gap.

B.4 TRUNCATED-POLL ROBUSTNESS

The masked draws above sample at temperature 1 over the full option support; self-consistency pipelines in practice truncate the sampling support (top-p nucleus, top-k). We re-run the count poll under truncated support computed from the same per-item option softmax masses (`results/asm_truncated_poll.json`; top-$p \in \{.9, .95\}$ and top-$k \in \{2, 3\}$, K=12, seeded): the one-forward advantage never reverses on any suite — ts v3 +.94pp at full support vs. +.53pp at top-p=.9 and +1.00pp at top-k=2; wt v2 +.55 vs. +.33/+.28; ts v2 +.66 vs. +.47/+.73 — and a single truncated draw remains decisively worst on every cell (top-p=.9 single-draw acc .884/.935/.906 vs. the poll .921/.950/.935). Truncation shrinks the poll's tail noise (so the gap narrows, as expected), but it does not reverse the ordering in favor of the count poll: the zero-draw mass read wins at the sampling-pipeline defaults as well as at temperature 1. *CSQA leg* (`results/asm_csqa_truncated_poll.json`, the 40–65% base-accuracy regime): the one-forward advantage again never reverses on the established suite — full support +.85pp (.660 logit vs. .651 poll), top-p=.9 +.96pp, top-p=.95 +.96pp, top-k=2 +.64pp, top-k=3 +.43pp — and a single truncated draw is the worst cell again (top-p=.9 single .614 vs. poll .650). At temperature 2 the draw-free advantage *widens* (full support +.63 → +3.18pp; top-p=.9 +3.51pp), because the higher temperature only inflates the poll's sampling noise and leaves the one-forward logit read unchanged. The dominance survives at base accuracy .66, with smaller absolute gaps at temperature 1 than on the near-tie synthetic suites, consistent with the finite-k identity.

Score	mass margin, 0 draws quintile ρ (1→5)	inv.	count margin, 12 draws quintile ρ (1→5)	inv.
TinyStories v3	1.00/.999/.970/.853/.569	0	.983/.987/.987/.879/.555	1
TinyStories v2	1.00/.999/.974/.899/.613	0	.983/.990/.990/.918/.604	1
WikiText v2	1.00/.997/.998/.959/.660	1	.992/.994/.997/.962/.669	2

Table 4: Quintile calibration of served per-call reliability (drawn from `results/asm_covgate_calib.json`). The zero-draw mass margin is monotone non-increasing (0 inversions on TinyStories); draw margins are not. At matched coverage the zero-draw serve also yields the most reliable served pool on every dataset.

B.5 Sampling required to approach exact-score decisions

We next quantify the cost of the draw-only alternative. For the answer, agreement of the k-draw count poll with the one-forward logit argmax is $96.7/97.5/98.6\%$ at k=12 (ts v3 / ts v2 / wt v2); fitting $1 - a\,e^{-\lambda k}$ gives $k_{99} \approx 19/14/12$ draws to reach 99% agreement — and in no measured cell does the poll exceed the exact logit answer. For the serve score, the count-margin deficit $g(k) = \mathrm{acc}(\delta(k)) - \mathrm{acc}(\delta^m)$ at c=.5 decays roughly as a power law $g(k) \propto k^{-p}$, $p \approx .76..96$, extrapolating to $k^* \approx 17/31/30$ draws to close the gap to $\leq$ .5pp (cov50; extrapolative).

B.6 Seed stability

A second independent draw seed (1001) shifts none of the headline numbers beyond $\sim$ .5pp on either dataset: ts v3 pooled answer .919 → .920, serve deficit (mass − count@12) @cov50 .013 → .012, $P(\hat{a}^L \neq \hat{a}(12))$ .033 → .030 (max $|\Delta|$ = .003); wt v2 .949 → .947, .005 → .008, .013 → .017 (max $|\Delta|$ = .004). The logit-based serve is seed-invariant by construction; the draw-free dominance holds under both seeds on both families. We do *not* claim draws are useless in a black-box setting with no logits. On these cloze suites a single forward matched or exceeded the repeated-sampling budget under both seeds; because computing the menu flag of Section 2 requires such a forward, this is the setting studied here (the established-menu exception for SmolLM2 is in Appendix B.9).

B.7 Calibration and risk–coverage comparisons

A serve score should be a *calibrated confidence*: served reliability should decrease monotonically as the score's quantile drops. Table 4 measures this for the mass margin [0 draws] versus the count margins [$4/6/12$ draws]: the zero-draw score is monotone non-increasing in quintile reliability on TinyStories (0 inversions) and nearly so on WikiText (one $+0.0007$ inversion), with quintile reliability spanning $1.0 \to .57$; every draw-based margin has at least one inversion on the TinyStories family (count@12: 1/1, count@4/6: 1/2 on ts v3 / ts v2). At matched coverage the mass margin also serves the *most reliable* pool on all three datasets (e.g. coverage 50% served per-call reliability ts v3 .997 vs. .985 for count@12, ts v2 .997 vs. .988, wt v2 .999 vs. .994). Caveat: the score is rank-calibrated (ordering), not a calibrated probability — the top quintile is a ceiling ($\rho \approx 1$); the calibrated quantity is the served-pool reliability reported here. Table 5 compares served accuracy of the zero-draw and draw-based selectors at 20/50/80% coverage, and Table 6 gives the served risk–coverage operating points.

Calibrated probability (isotonic, out-of-sample). Isotonic (PAVA) regression of the real per-call reliability on the serve score, fit on a seeded half and evaluated on the other half, yields a deployable probability. The zero-draw mass margin calibrates at least as well as the draw-based scores: out-of-sample ECE (10 equal-count bins) is $.014/.021/.009$ on ts v3 / ts v2 / wt v2 (vs. $.016/.026/.016$ for the 12-draw margin and $.053/.045/.032$ for the 6-draw margin), with comparable calibration MSE. The result is *targetable*: serving items whose calibrated $\hat{\rho} \geq .95$ realizes served per-call reliability $.995/.991/.992$ (coverage $50\%/56\%/78\%$), and $\geq .99$ realizes $.995/.996/.999$ (coverage $49\%/43\%/59\%$) out of sample. Thus, with an out-of-sample isotonic fit, the zero-draw score also yields a calibrated *probability* on these suites, not only an ordering.

Not transportable across menu states. We test whether the calibrated probability remains valid after re-menuing, or whether each menu variant must carry its own fit-half. Measured on the two suites

Serve selector	ts v3	ts v2	wt v2	row note
mass margin (0 draws)	1.00/.999/.976	1.00/.999/.985	1.00/.999/.994	acc @ 20/50/80
top-1 confidence (0 draws)	1.00/1.00/.977	1.00/.999/.984	1.00/.999/.994	
entropy (0 draws)	1.00/.999/.978	1.00/.999/.984	1.00/.999/.994	
count margin K=4	.957/.964/.960	.953/.968/.970	.981/.983/.983	
count margin K=6	.980/.977/.968	.970/.979/.975	.978/.987/.988	
count margin K=12	.983/.985/.974	.983/.988/.981	.992/.994/.988	
random floor (@50%)	.917	.939	.949	(5 seeds, sd $\leq$ 4.6e-3)

Table 5: Served accuracy at coverage 20/50/80 by selector. The three zero-draw selectors are mutually nearly identical and strictly dominate all draw-based margins; draw margins converge to the free value monotonically in k and never exceed it. 1000-resample nonparametric 95% CI of (mass margin − count@12): at $c = .5$ positive on all three datasets ($[.005, .021]$ ts v3, $[.003, .020]$ ts v2, $[0, .010]$ wt v2); at $c = .8 \approx 0$ (CIs straddle zero). Source: `results/asm_covgate_baselines.json`.

with real re-menu draws (`results/asm_calib_transport.json`): observed-menu ECE is $.011/.009$ out of sample, but the *re-menu* realized reliability at the same deciles is systematically lower by $\bar{\Delta} = .246/.329$ and the ECE against the re-menu outcome is $.250/.328$ — the observed calibration is accurate for the observed menu and *not valid for the served-into menu*. Consistency check: this is why policy F applies the per-bin deflation factor η (fit-half r_c/r_w) to the calibrated score rather than re-fitting $\hat{\rho}$ on the re-menu rows: the menu state changes the level but not the within-menu ordering, so a single transport term (the $\eta(\text{flag}, \text{bin})$ of Section 6) recovers calibration across states on these two cloze suites, without a separate isotonic fit for each menu variant. On single-token CommonsenseQA the benefit of the same deflation is checkpoint-dependent (Section 6, Table 3).

dataset	serve	risk@50	risk@90	max cov @ risk $\leq 1\%$	$\leq 2\%$	$\leq 5\%$
ts v3	mass (0 dr)	.001	.039	65%	78%	94%
	count@12	.015	.047	17%	74%	92%
ts v2	mass	.001	.027	74%	87%	97%
	count@12	.012	.033	13%	80%	97%
wt v2	mass	.001	.017	85%	92%	100%
	count@12	.006	.022	79%	89%	100%

Table 6: Served risk-coverage and operating points (for the comparison of Section 2). "max cov @ risk" = the largest coverage whose *cumulative average* served risk stays within the budget (definition in `results/asm_riskcov.json`, together with bootstrap CIs at cov 50/90).

B.8 Risk-controlling baselines and their limitations

Head-to-head with risk-controlling/conformal serving baselines. We evaluate the deployable protocol (F) on the suites where the flag helps against the model-free "risk-controlling" selective-prediction baselines cited in Section 7 (`results/asm_rcps_serve_e2e.json`; same per-item arrays and OOS split; every rule selects its threshold on the fit half and is evaluated on the eval half at the fitted served count). We compare three threshold rules at covered-risk targets $\alpha \in \{0.05, 0.1, 0.2\}$: RCPS (Bates et al., Hoeffding upper bound on the mean per-item covered risk $1 - \rho^{\text{remenu}}$ over the served prefix, confidence $\delta = .05$); the empirical EMP threshold used in the serving analysis, on the same deflated score (fit-half *empirical* prefix mean, no margin); and EMP-MARGIN, the same empirical rule on the raw margin (the un-gated conformal reference). *Finding 1: on this K=12 granularity and ∼900-item fit half, the Hoeffding margin is too large for small budgets —* RCPS *serves at* none *of $\alpha \leq .1$ (ts v3 and wt v2), appearing only at $\alpha = .2$, where its eval risk is $\leq \alpha$ (ts v3 .126 at 57% coverage; wt v2 .155 at 63%) but at a lower coverage than the empirical rules. Finding 2: the empirical thresholds over-select on the fit half where a risk guarantee would constrain selection: at $\alpha = .1$,* EMP *on ts v3 reaches 38% coverage at realized eval risk .097 ($\leq .1$), and at $\alpha = .2$ realizes $.174/.211$ (ts v3 / wt v2) on 69% coverage — the wt v2 cell*

misses α by 1.1*pp, the same over-transfer of item-block noise shown by the fit-5% cells of Table 15.* EMP-MARGIN (flag-blind) realizes the *worst* served value at matching coverage (.176/.214 at ts v3 38%/wt v2 43%), i.e. the deflation step is worth $\geq$ .07 of covered risk at the same served coverage — not a renormalization artifact, a measured serve-level difference. *Finding 3: the only rule that misses its budget is the least conservative one, and it misses by the small-prefix noise of those cells.* The comparison is therefore informative rather than adversarial: a model-free guarantee (RCPS) is too expensive at $K{=}12$ item granularity for the $1-10\%$ budgets studied here; the empirical rule carries no distribution-free promise and instead requires reporting that over-transfer; and both beat the un-gated conformal reference at the same coverage.

B.9 Answer and serving comparisons on CSQA across four checkpoints

The synthetic draw-dominance analysis (Appendices B.7 and B.5) is extended unchanged to the established, human-distractor CSQA-941 menu for all four generators: real $K{=}12$ masked draws (fp32, seed 20260903, temperature-matched), menu probabilities from each family's own length-normalized likelihood artifact — the established-menu analogue of the single-forward 5-option logit probe — and paired bootstrap CIs (2000 resamples, one resample set shared by both selectors) in `results/asm_csqa941_draw_calib_4fam.json`, Table 7. (i) *The finite-k direction replicates with significance on all four families*: count@12 minus one draw (pp, [95% CI]) is $+4.3[2.1, 6.6]$ (g1), $+8.1[5.4, 10.8]$ (g2), $+4.9[2.1, 7.5]$ (SmolLM2), $+2.0[0.3, 3.7]$ (Qwen3-4B) — every CI excludes 0, the established-menu counterpart of the count-poll law (Appendix A.1). (ii) *Answer-level parity holds on three families; the fourth is a bounded, reported violation.* Logit minus count@12 (pp, [CI]) is $+1.2[-0.2, 2.7]$ (g1), $-0.1[-2.1, 1.9]$ (g2), $-0.1[-1.0, 0.9]$ (Qwen3-4B) — indistinguishable from parity — while SmolLM2 gives $-4.2[-6.5, -1.9]$: the twelve-draw poll answers about four points *better*, the single measured cell where draws add answer accuracy, quoted with its CI rather than suppressed; the same family is the depth-outlier of the abstention map (Table 18). (iii) *Serve level: the synthetic strict dominance replicates at* 20% *coverage for the three moderate families and significantly* reverses *for* SmolLM2*; at* 50/80% *coverage the two selectors are statistically indistinguishable (on the reference draws every paired CI includes or touches* 0*; the three-draw-seed analysis follows below)* (deficit CIs in Table 7). *The reversal is localized to the cov*20 *head pool.* Decomposing that pool into its two-way displacement between the mass and count orders (89 items each way) accounts for the whole deficit: the mass order pulls in items on which the logit answer is right .438 of the time vs. .607 for their count-order counterparts, and these are precisely the deep-tail items where the probe is over-confident against the draws (off-menu threat 900 vs. 313; probe margin .970 against draw margin .489; .180 with zero gold frequency in the twelve draws). At cov 50/80 the displacement Δ of logit correctness flips positive (+0.9/+4.1pp), i.e. the two selectors disagree materially only when the order concentrates service into the over-confident tail (decomposition, `results/asm_csqa941_smol_reversal_decomp.json`; `scripts/asm_csqa941_smol_reversal_decomp.py`). By contrast, the answer-side gap does not share this deep-tail signature: the 85 items right only under the twelve-draw vote are threat-typical (threat median 39.3 vs the population median 37.4; probe-margin median .613 $\approx$ .622), while the 46 items right only under the logit answer sit in the draw-indeterminate band instead of the tail (draw-margin median .125 vs the population .583; .239 of them at zero gold frequency among the draws) — the parity violation is spread across the menu rather than concentrated in an over-confident pocket, so it remains what (ii) reports: one bounded exception, quoted with its CI. *Multi-seed: every signature above is stable across two further draw seeds* (20260904/20260905; same data, runner, $K{=}12$ and protocol; draws in `results/raw/asm_qwen15_csqa941_{g1,g2,smol,q4b}_s{04,05}.json`). Count@12 beats single on every family on every seed (range of the delta over the three seeds: g1 $[.039, .047]$, g2 $[.022, .081]$, SmolLM2 $[.037, .049]$, Qwen3-4B $[.018, .023]$); the logit-vs-count gap keeps each family's sign class (g1/g2/Qwen3-4B inside $\pm.014$: $[.002, .012]$/$[-.001, .014]$/$[-.006, -.001]$; SmolLM2 negative on all three seeds, $[-.046, -.021]$); and the cov20 reversal replicates twice ($-.053/-.074$) with 91/86 items displaced each way and the same logit-correctness gap direction (.472/.419 vs .582/.581); set identity separates the two mechanisms — the mass-order side of the displacement is seed-stable (Jaccard .59/.68/.62; the over-confident pocket is the same items, since probe margins and threat are draw-free) while the items selected by the count order change with the draws (.26/.25/.25; $K{=}12$ counts churn the boundary), so what replicates is the gap's direction and size, not an item list (displacement ids and overlaps: `results/asm_csqa941_smol_reversal_`

jaccard.json; scripts/asm_csqa941_smol_reversal_jaccard.py; point estimates per seed: results/asm_csqa941_multiseed_parity.json; analysis script: scripts/asm_csqa941_multiseed_parity.py). Serve-level pairing CIs locate precisely where each strength lies (same 2000-draw paired bootstrap as the reference cell): the cov20 dominance is strict on all three seeds for g1 and Qwen3-4B ($3/3/3/3$) and strict only on the reference seed for g2 ($1/3$; the other two cov20 CIs touch or include zero), while the SMOLLM2 reversal is strict-negative on two seeds ($2/3$), its 20260904 CI straddling zero at $[-.112, .027]$; at cov50/80 the include-or-touch-zero reading holds on $22/24$ seed-by-coverage cells and the remaining 2 are both upward (20260904, cov50: g1, g2) — no draw seed shows a harmful separation anywhere (per-seed CIs: results/asm_csqa941_multiseed_deficit_cis.json). *View basis:* the chain flags each item on the masked first-token view of its option masses (identical to the strict length-normalized view wherever all five options are single tokens; only SMOLLM2 carries 156 two-token-option items, of which the 820-item join keeps 35). Swapping the flags to the strict length-normalized view on the identical script and draws (results/asm_csqa941_basis_sensitivity.json) moves the SMOLLM2 flag set by 6 of 820 items ($0.946 \to 0.939$) and no conclusion class: the F−B cov50 CI still includes zero ($+0.011[-0.004, +0.024] \to +0.013[-0.003, +0.028]$); the same swap on the g1 leg is exact to rounding (flag inventory unchanged; risks move $\leq .0003$). Quintile monotonicity of the mass-margin rho profile holds for g1 and Qwen3-4B and fails for g2 and SMOLLM2, echoing the weaker families' inversions on the synthetic suites. Scope: this is one suite of 941 items; sub-menu $M \in \{2, 3, 4\}$ cells are measured below (Appendix D, threat near M-invariant); architectures beyond these four checkpoints are not tested.

Family	Single/count@12/logit accuracy	$P(c \neq L)$
g_1 Qwen2.5-1.5B	.605/.648/.660	.068
g_2 Qwen3-0.6B	.377/.458/.457	.163
SMOLLM2-1.7B	.473/.522/.480	.215
Qwen3-4B	.634/.655/.654	.039

Family	Count−single pp [95% CI]	Logit−count pp [95% CI]	Mass−count pp @20/50/80
g_1 Qwen2.5-1.5B	$+4.3[2.1, 6.6]$	$+1.2[-0.2, 2.7]$	$+11.2/+1.1/+0.4$
g_2 Qwen3-0.6B	$+8.1[5.4, 10.8]$	$-0.1[-2.1, 1.9]$	$+4.8/+1.9/+0.5$
SMOLLM2-1.7B	$+4.9[2.1, 7.5]$	$-4.2[-6.5, -1.9]$	$-8.0/+0.2/+0.5$
Qwen3-4B	$+2.0[0.3, 3.7]$	$-0.1[-1.0, 0.9]$	$+17.0/+1.1/-0.4$

Table 7: Draw-dominance summary on the established CSQA-941 menu (real $K=12$ masked draws, fp32; menu probabilities from each family's own 5-option likelihood softmax; source results/asm_csqa941_draw_calib_4fam.json). The answer columns quote accuracy of the single draw, the 12-draw count argmax, and the zero-draw likelihood argmax, with paired 95% bootstrap CIs (2000 shared resamples) for both deltas; $P(c \neq L)$ is the probability that 12 draws change the answer selected by the single forward pass. The rightmost column shows mass−count@12 serve deficits at coverage $20/50/80$; paired CIs: g1 $[4.8, 14.9]/[-0.9, 3.4]/[-1.2, 1.7]$, g2 $[0.5, 11.2]/[0.0, 5.1]/[-1.9, 1.6]$, SMOLLM2 $[-14.9, -0.5]/[-3.2, 2.8]/[-0.9, 2.4]$, Qwen3-4B $[8.0, 19.7]/[-1.5, 3.2]/[-1.3, 1.2]$ (pp, same bootstrap).

B.10 ANSWER ACCURACY WITH MULTI-TOKEN SCORES

A natural concern is that the draw-free evidence rests on $M=5$ single-token cloze. We therefore run the answer-layer comparison on the two *multi-token* established suites that we already score — reading each item from its likelihood-conditional distribution $p \sim \mathrm{softmax}(\mathrm{opt_loglik})$, drawing $K=12$ i.i.d. reads, and comparing the one-forward argmax reader to the count poll (results/asm_multitoken_draws_consume.json): on full CSQA ($M=5$, $N=1197$) the logit argmax is .610 vs. count poll .587 (+2.3pp; single draw .553, $P(\text{disagree}) = .107$, and among disagreements the logit read is correct 48%), and on ARC ($M=4$, $N=1460$) .469 vs. .414 (+5.5pp; single draw .375, $P(\text{disagree}) = .284$). Sampling noise among disagreements never favours the poll on these bases either, and the single draw is worst on both — the dominance gap is *larger* on ARC, the hardest items the suite exposes. This is a likelihood-basis cell (scope: no new masked model for-

ward passes), but it moves the answer-layer claim off the single-token anchor to multi-token, $M \neq 5$ menus with human distractors, the transfer direction the evidence supports.

B.11 Cross-task numerical summary

Table 8 collects the controlled cloze results of this appendix, and Table 9 gives whole-set answer accuracy by answer source.

		ts v3	ts v2	wt v2
Answer acc	logit (1 fwd)	.927	.939	.953
	count@12 (12 draws)	.919	.937	.949
Serve acc @50	mass margin (0 dr)	.999	.999	.999
	count@12	.985	.988	.994
CI(mass−count@12) @50		[.005, .021]	[.003, .020]	[.000, .010]
Calib ECE (isotonic)	mass (0 dr)	.014	.021	.009
	count@12	.016	.026	.016
Menu flag	$P(w^* \mid \tau \geq 1)$	.942	—	.915
	$P(w^* \mid \tau < 1)$	.095	—	.059

Table 8: Unified cross-task view (source: `results/asm_menu_cross_uni.json`). The draw-free serve dominates or matches on every dataset; the isotonic zero-draw probability calibrates at least as well as the paid score; the gold-free menu flag is measured on TinyStories v3 and WikiText v2 (ts v2 not instrumented for the flag).

Answer source	TinyStories v3	TinyStories v2	WikiText v2
single draw ($K=1$)	.872	.886	.927
logit argmax (1 forward, 0 draws)	**.927**	**.939**	**.953**
pooled count argmax ($K=12$)	.919	.937	.949
$P(\hat{a}^L \neq \hat{a}(12))$	.033	.024	.013
among disagreements: logit-only / counts-only	.55 / .33	.50 / .39	.65 / .30

Table 9: Whole-set answer accuracy by source (real draws; gold only in evaluation) and the answer-source discrepancy. In these three disagreement cells, the exact-score answer is correct more often than the count answer. This is an empirical comparison against gold, not a consequence of convergence to the score argmax. Sources: `results/asm_covgate_baselines.json`, `results/asm_covgate_discrep.json`.

C Off-menu candidates and sensitivity to menu construction

C.1 Characterizing counterexamples to the flag

The flag's takeover rate on the cloze suites ($P(w^*$ becomes the pooled answer $\mid \tau \geq 1) \approx .92$–$.94$; Table 8) is a *plurality* statement, and part of it is the ordering algebra: inserting the mass-preferred excluded candidate and having it win pluralities is, in the large-τ regime, close to re-reading the same logits. This subsection analyzes the residual — the 5–9% of flagged items on which w^* does *not* become the pooled answer (real $K=12$ re-menu draws, all three sets; `results/asm_flag_counterexample.json`). Three facts separate the flag from the ordering restatement:

The residual consists of plurality-stable items rather than counter-preference flips. Using the strict content reading (the re-menu pool answer is the w^* string), w^* takes the plurality on $.926/.908/.915$ of flagged items (ts v3 / ts v3-g2 / wt v2); the 7–9% residual decomposes into only $\approx$ .3–.7 points of third-string flips (where the re-menu pool answer changes to neither the original answer nor w^*) and $\approx$ 7–9 points of *plurality-stable* items (the original pooled answer survives). Auxiliary flips (answer changes under the re-menu) still concentrate on flagged items ($P(\text{flip} \mid \tau \geq 1) = .93/.91/.92$ vs. $\leq .06$ clean), so the residual remains consistent with the flag's menu-dependence claim.

The counterexamples sit at the flag's own threshold. Mean $\log \tau$ is 0.30–0.41 on plurality-stable counterexamples versus 2.1–2.5 where w^* takes over, and the counterexample rate falls monotonically with τ (no-change rate among flagged: $0.28/0.42/0.49$ in the threshold band $\tau \in [1, 1.4)$, near 0 for $\tau \geq 3$, where w^* wins essentially always). I.e. the counterexamples are the items where the excluded leader is only marginally ahead of the presented leader, so the $K{=}12$ plurality is decided by draw noise; the ordering law is not violated, it is just not a deterministic predictor of the finite-sample plurality.

The flag still detects loss of gold-draw reliability. The plurality-stable counterexamples lose a substantial proportion of the gold's per-call reliability on the altered menu: realized drop 0.36 (ts v3; 30/30 items with $\Delta^{\text{real}} > 0$), 0.26 (ts v3-g2; 40/49, exact sign-test $p = 5 \times 10^{-6}$), 0.31 (wt v2; 46/55) — 2.4–3.6× the clean drop (0.10–0.11) and close to the drop where w^* actually flips. On ts v3 these same items carry a logit-space expected drop $\Delta = 0.49$ (the real-draw runner; vs. 0.64 where w^* flips and 0.10 clean). So the flag's informational content is a *distributional* menu-hostage law — broadening the menu substantially cuts the gold's expected per-call reliability even when it does not switch the plurality — which no within-menu first-order statistic restates (Appendix C.3 compares τ with within-menu statistics directly). The near-tautological reading holds only for the large-τ projection of this law.

C.2 Menu hardness: sensitivity to adversarial distractor strength

A natural concern is whether the draw-free and threat results are artifacts of easy, well-separated menus. We address this concern in two ways. First, the construction (Section 2) already uses context-plausible near-tie distractors (a surrogate conditional-bigram reader gives the gold $q \in [.45, .63]$ of the conditional mass on the constructed menus, solving 0.906 of them vs. chance .2). Second, we sweep the distractor set to the *adversarial* extreme within the generator's own logit space. The extended-S masses are a real softmax over the menu plus the generator's top-16 excluded continuations, and the 5-subsets conditionals (masked softmax is ratio-invariant) give, for any menu, the gold's conditional mass q_g (expected single-draw reliability), the deterministic argmax, and the same-basis off-menu threat τ. For each item we build $k = 0, \ldots, 4$ menus $\{g\} \cup \{\text{top-}k \text{ excluded}\} \cup \{\text{top } (4{-}k) \text{ distractors}\}$ and measure the curve (Table 10; `results/asm_menu_hardness.json`).

The construction is *validated, not assumed*, in three places: (i) at $k{=}0$ it reproduces the observed argmax accuracy and flag rate on every set (deviation $\leq$ 0.13pp); (ii) the $k{=}1$ menu equals the real-draw re-menu, and the computed conditional reproduces the *measured* p^{remenu} (mean $|d| = 6{\times}10^{-4}$, Spearman 0.9996, $N{=}1481$), so the whole curve inherits the real-draw confirmation of Section 2; (iii) the $k{=}0$ expected reliability matches the logit artifacts' p_g (e.g. wt v2 0.9227 vs. 0.9224). With the mask unchanged, each higher-k row is the same deterministic forward restricted to a harder menu — the only thing the row adds is *distractor strength*, which is the axis in question. *Real-draw confirmation* ($K{=}12$ draws, same protocol): the extended-S predictions at $k{=}2$ and $k{=}4$ are confirmed on real draws over the literal adversarial menus ($N{=}1467/1456$; gold at options[0]). Realized gold per-call reliability 0.594 vs. predicted $q_g = 0.593$ at $k{=}2$ and 0.615 vs. 0.613 at $k{=}4$ (0.1–0.2pp; per-item Spearman 0.96– 0.97). On the same items the observed-menu realized reliability is 0.879, so the adversarial swaps reduce the gold's realized per-call reliability by 0.285 ($k{=}2$) and 0.264 ($k{=}4$) on average, matching the predicted $0.285/0.265$: menu hardening removes roughly a third of the gold's expected reliability in real sampling. On the same item subsets the realized 12-draw poll answer stays *below* the one-forward logit answer at every menu: on the $k{=}2$ subset, 0.920 vs. 0.928 on the observed menus and 0.655 vs. 0.661 on the adversarial menus; on the $k{=}4$ subset, 0.920 vs. 0.929 (observed) and 0.671 vs. 0.686 (adversarial). The poll deficit is largest (1.5pp) at the fully adversarial menu: the draw-free dominance is not an artifact of easy, well-separated menus — the count poll is *most* dominated where the menu packs near-tie strong candidates.

Reading. (a) The observed regime sits at the easy end, but not because the distractors are weak: context-plausible near-ties already put $\bar{q}_g$ at .83–.92 and the deterministic answer at .89–.95. (b) A *single* candidate swap to the generator's strongest excluded preference — the re-menu construction — cuts expected gold reliability by $\Delta{=}.26/.28/.32$ on the three sets (matching the measured intervention), and makes the deterministic one-forward answer *the excluded candidate* on 29–35% of

distractor regime	$\bar{q}_g$ ts v3 (g1/g2)	$\bar{q}_g$ wt v2	argmax acc (wt)	w*-flip (wt)	argmax is g (w*) g1	argmax is g (w*) wt
k=0 observed	.879 / .826	.923	.954	.000	.929 (-)	.954 (-)
k=1 +w*	.620 / .547	.599	.619	.367	.661 (.300)	.619 (.367)
k=2	.594 / .517	.574	.619	.367	.661 (.300)	.619 (.367)
k=3	.587 / .508	.565	.619	.367	.661 (.300)	.619 (.367)
k=4 fully adversarial	.613 / .533	.573	.627	.373	.686 (.314)	.627 (.373)

Table 10: Menu-hardness curve (logit space; gold in evaluation only). $\bar{q}_g$ = expected gold reliability (single-draw expected accuracy); "argmax (wt)" = deterministic one-forward answer accuracy on wt v2; "w*-flip" = fraction where the deterministic answer becomes the *excluded* candidate after the menu change; "argmax is g (w*)" gives ts v3 g1 accuracy with the w*-flip parenthesized. Each k replaces k weakest distractors by the generator's top-k *excluded* preferences. The k=1 row is the real-draw intervention and matches its measured value. A dash marks k=0, where no w^* is on the menu.

items that were clean under the observed menu. In other words the dominance claims are menu-relative by construction: they hold for the observed composition, and fail as soon as the menu under-covers the generator's preference — the exact situation the flag exists to catch. (c) The fully adversarial k=4 menu (gold plus the four strongest excluded continuations) still has the gold as the one-forward argmax on ≈two-thirds of items, but its expected reliability drops from $\sim$ .9 to $\sim$ .55; the off-menu threat itself collapses ($\tau \ll 1$) because the strongest candidates are no longer excluded. Two distinct failure regimes are therefore visible — *under-inclusion* (flag large, flagged answers hostage, the large-τ law) and *within-menu confusion* (all strong candidates present, reliability low, threat silent): the flag measures which menu state the answer is in. (d) Under the observed (random/contextual) construction, the generator's strongest excluded preference exceeds the weakest *included* distractor on 99.5% of items — random menu composition is drastically non-greedy, and the menu-incompleteness the flag reads is a property of the menus, not an artifact of one benchmark. (e) *A calibrated identity bounds the practical stakes: the flag's call rate equals the menu-determined fraction.* Per item on the observed nearby-tie menus we compare two real-draw events — $\tau_i \geq 1$ on the observed menu vs. the K=12 count-argmax on the literal one-candidate re-menu returning the excluded candidate w_i^* (real re-menu draws; `results/asm_flag_menudeterm_identity.json`): the sets nearly coincide, $P(\tau \geq 1) = .301/.372$ vs. $P(w^*\text{-flip}) = .309/.376$ on ts v3 / wt v2 ($|\Delta| \leq .007$), with $P(w^*\text{-flip} \mid \tau \geq 1) = .93/.92$ and $P(\tau \geq 1 \mid w^*\text{-flip}) = .91/.91$. On a near-tie menu the answer is set by menu composition on the flagged items and, in the measured cells, only there. Consequently *quality-matched (greedy) menus absorb the mechanism*: once even one distractor is the generator's own strongest excluded candidate the flag rate collapses (.30 → .003, rows $k \geq 1$) because the strongest continuation is already presented and nothing latent can displace the leader — the composition-hostage set is empty by construction, i.e. Reading (c)'s second regime. The two practical readings are complementary: on a fixed/human candidate profile (CSQA, $P(\tau \geq 1) \approx .65$) the flag's call rate is literally the fraction of answers controlled by the menu composition; on a self-generated candidate profile, building the menu from the generator's own top preferences removes the mechanism, at the cost of the trimmed difficulty ($q_g \approx .59$–$.61$ above).

C.3 Ablation of the flag and sensitivity to the construction rule

We compare the off-menu statistic $\tau = p^{\text{off}}_{\max}/p^{\text{lead}}$ against its own ingredients and simpler within-menu competitors for the two jobs the flag performs (predicting pool-answer displacement and ordering the re-menu gold-reliability drop). Table 11 does that on the real draws of both suites. We report six observations; budget and lead sensitivity follow in Appendix C.4.

(a) The threat family separates from the within-menu family. On both jobs and both suites, τ, its numerator $p^{\text{off}}_{\max}$ and its denominator p^{lead} all score far above the within-menu statistics: displacement AUC 0.95–0.99 vs. margin 0.71–0.75 (extended-S margin and entropy intermediate, 0.80); drop-ordering Spearman 0.74–0.85 vs. margin 0.18–0.32. The orthogonality claim of Section C.1 is therefore not just a partial-correlation residual: as stand-alone order statistics, menu-external mass dominates menu-internal scores.

(b) Head-to-head: τ is not better than its numerator. Normalization by p^{lead} adds nothing for displacement (τ AUC .971/.986 vs. p^{off} .947/.961) or for the drop-ordering job (Spearman .786/.847 vs. .793/.840); $-p^{\text{lead}}$ alone is comparable (AUC .975/.983). The distinguishing information for both jobs is the *extended-S off-leader mass itself*; the ratio form was chosen for its threshold scale ($\tau \geq 1$), not because the ordering requires it.

(c) The ratio's threshold transfers across suites; the mass threshold does not. Cross-suite threshold-transfer (`results/asm_tau_p_off_transfer.json`; ts v3 / wt v2 / CSQA single-token, all extended-S off-menu artifacts on Qwen2.5-1.5B): $\tau \geq 1$ is a scale-free condition ("off-candidate stronger than the presented leader") whose call rate adapts to each suite automatically, $P(\tau \geq 1) =$ .307/.372/.721; a *frozen* absolute mass threshold (anchor: CSQA p^{off} median $=$ .256, the deep-under-inclusion suite) over-flags the synthetic suites when transferred ($P(p^{\text{off}} \geq c) =$ .367/.425 vs. .307/.372): the mass rule would need a per-suite retuned cut (p^{off} threshold to reproduce the ratio call rate: .308/.300 on ts v3 / wt v2, .183 on CSQA) — i.e. a $\sim 1.7\times$ downward drift from the frozen cut. These two findings are therefore complementary: within a suite τ and p^{off} carry the same ordering information (b); across suites the ratio's unit threshold is the transferable object, because the two measures share a rank law under menu renormalization while absolute off mass is compressed by deep under-inclusion (CSQA) and dilated on the firmly-covered synthetic cloze items. *Caveat:* the off-candidate construction differs across suites (bigram-filtered content words on synthetic, generator-native top-k on CSQA), so the transfer statement is about the threshold *scale* being stable, not about the candidate sets being identical.

(d) A fourth suite with a distinct construction. ARC-Challenge adds a $M{=}4$, multi-token, generator-native (log-likelihood) basis (`results/asm_tau_p_off_transfer_arc.json`): its under-inclusion call rate $P(\tau \geq 1) =$ .333 (validation .393 / test .318) sits with the synthetic-cloze family (ts v3 .307 / wt v2 .372), i.e. the *unit* ratio threshold needs no retuning on a fourth suite built differently (length-normalized MC likelihood ratios rather than the extended-S mass softmax), while the suite's absolute off-mass logit scale ($\text{med} \log p^{\text{off}} \approx -27.6$) is more than an order of magnitude below the mass suites' median logit (-1.87 ts v3 / -1.68 wt v2 / -1.36 CSQA): an absolute cutoff carried over would be meaningless. This is the measured sense in which "$\tau \geq 1$ transfers without per-suite retuning" is a claim about the threshold scale, not about any single construction's candidate set.

(e) The drop-ordering is menu-residual, not a probe. We ask how much of the flag's ordering information is a re-statement of within-menu difficulty (a "probe" any selective-prediction baseline could read) versus information about the menu outcome (`results/asm_probe_menu_decomp.json`). We residualize $\log \tau$ against the within-menu probe (logit margin plus calibrated $\hat{\rho}_w$, fit-half, OOS evaluation) and re-rank the residual against the realized drop $\Delta = \rho_w - \rho_c$: removing the probe's linear prediction leaves the drop-ordering *intact* (Spearman .775 $\rightarrow$ .793 ts v3; .846 $\rightarrow$ .846 wt v2) while the probe itself is negatively associated with the drop ($-.18/-.31$). The flag's ordering content therefore lives entirely on the menu-outcome side; within-menu difficulty *explains none* of it (and correlates with it negatively). Covered-risk @50 is .57/.67 for the residualized score versus .63/.72 for the raw flag on ts v3 / wt v2.

(f) Sensitivity to the re-menu construction rule. The intervention law's construction replaces the *weakest* non-gold distractor with w^*. Recomputing the predicted drop under two alternatives (drop the *strongest* distractor, or a random one) on both families (`results/asm_remenu_rule_sensitivity.json`) shows the mean is only mildly rule-sensitive ($\Delta =$ .260/.226/.250 Qwen and .214/.179/.204 SmolLM2 for drop-weakest/-strongest/random), and the per-item ordering is stable (Spearman $\geq$ .90). The strictly monotone form of the law, however, is specific to our drop-weakest rule: under "drop the strongest" the two lowest-τ quintiles turn *negative* ($-.033/-.022$ for Qwen, $-.026/-.034$ for SmolLM2) and $P(\Delta > 0)$ falls to .76–.81 — removing a strong in-menu distractor helps the gold more than the weak w^* insertion hurts it. The high-τ content is rule-independent (top quintiles .62–.70 across all rules). The construction choice is therefore principled (drop the least-mass distractor so the inserted w^* has maximal relative effect), and the law's *high-threat* claim is robust to it, while low-τ behavior is the most construction-sensitive part.

score (riskier→high)	AUC (displacement) ts v3	AUC (displacement) wt v2	Spearman (drop) ts v3	Spearman (drop) wt v2
$\tau = p^{\text{off}}_{\max}/p^{\text{lead}}$	.971	.986	.786	.847
$p^{\text{off}}_{\max}$ (numerator)	.947	.961	.793	.840
$-p^{\text{lead}}$ (denominator)	.975	.983	.738	.829
− within-menu margin	.711	.746	.181	.317
− ext.S margin	.799	.802	.542	.594
ext.S entropy	.802	.819	.520	.600

Table 11: Flag ablation on real draws (`results/asm_flag_ablation.json`). Displacement AUC = discrimination of served content-level flip; drop Spearman = rank correlation of the re-menu gold-reliability drop $\Delta\rho$. The threat family dominates the within-menu family on both jobs; the ratio τ is not better than its numerator $p^{\text{off}}_{\max}$ (see text).

C.4 Sampling-budget and menu-leader sensitivity

The served flag law is robust to the draw budget: computing the pool answer of the *first* K real draws, $P(\text{flip} \mid \tau \geq 1)$ is already .857/.880 at K=4, .890/.888 at K=6, .906/.889 at K=8 and .940/.922 at K=12 (ts v3 / wt v2): the flag law holds at K=4 and only sharpens with sampling. A remaining worry about the p^{lead} denominator — precision degrading where the lead option is not gold ($\sim$ 8/5% of items) — does not materialize: within the lead-not-gold group the flag's displacement AUC stays $\geq$ .93, and $P(\text{flip} \mid \tau \geq 1)$ is .967/.985 (*higher* than the lead-correct group), so no proxy-conditioning of deployment τ is called for on these data.

C.5 Candidate-set sensitivity on the synthetic suite

Before moving to CSQA we re-run the ts v3 off-menu profile in the *generator-native* construction (`results/asm_offgen_tsv3_consume.json`; top-4096 `good`-filtered next-token candidates, identical to the CSQA recipe; N= 1479): the generator-native candidate set *raises* the under-inclusion call rate ($P(\tau \geq 1) = .389$ vs. .301 under the corpus-bigram profile), confirming the bigram set under-lists some candidates — but the additional generator-native flags are *weakly displaced* items ($P(\text{flip} \mid \tau \geq 1) = .530$ and a realized ρ-drop .392 for the generator-native flag vs. .933/.622 for the bigram flag), and the two τ orderings correlate only $\rho = .47$. So the corpus-bigram profile does not lose *informativeness by flagging fewer items*: it sees a smaller, sharper subset of the menu-hostage items, and the generator-native superset dilutes the flag's precision as the CSQA plausibility audit (Appendix E.7) predicted (shallow continuations). *Profile-robust flag.* The two profiles bound the deployment choice a practitioner faces (flag precision .93 on the curated-bigram profile vs. .53 on the generator-native superset). Decomposed per item on the real draws (`results/asm_flag_profile_intersection.json`), the disagreement is *one-directional in informativeness*: items flagged by *both* profiles retain the curated precision ($P(\text{flip} \mid \text{both}) = .924$, mean realized ρ-drop .62, call rate .214; "flip" here is the *content-level* event — the pooled answer string changes under the one-candidate supplement, the same event the ablation and the counterexample sections use; the w^*-specific takeover event of Section 2 is a separate, more restrictive reading); items flagged only by the bigram profile stay sharp (.953); and the generator-native-*exclusive* flags are the *least* threatened items on the suite ($P(\text{flip} \mid \text{only-generator}) = .047$, *below* the clean baseline .071), so the union only dilutes precision (.61). The profile-dependence of the flag is therefore measured: a practitioner who cannot commit to a candidate profile can require both profiles to fire — the intersection reproduces the curated precision at $\approx 71\%$ of its call rate, and the generator-native additions contribute near-zero-information items, not missed threats. The transfer below therefore pairs the generator-native construction with the bigram suite's sharper flag.

C.6 Draw-count (K) sensitivity of the law and the served-side quantum

We ask whether the re-menu law depends on K on the served side — the $1/K$ risk quantum caps tight budgets — and whether the predicted Δ still matches realized within noise at K=50. The stored draw paths are i.i.d. counts from each item's one-forward multinomial, so a K=50 realization is drawn from the item's empirical multinomial (plug-in bootstrap; `results/asm_`

remenu_K_sensitivity.json). The law is *invariant in expectation*: at $K{=}50$ the mean realized drop, the Δ-vs-$\log\tau$ ordering, the monotone quintiles, and the predicted-vs-realized fit are unchanged on both suites (mean .258 $\rightarrow$ .258 ts v3 / .323 $\rightarrow$ .323 wt v2; Spearman $=$ .786/.847; $|d| = .0815$, Spearman .85), and finer K only concentrates the realized noise around the same expectation ($P(\Delta_{\text{real}} > 0)$: .624 $\rightarrow$ .644 ts v3): the law's agreement with realized outcomes is not a $K{=}12$ artifact. On the served side the $1/K$ quantum is real but arithmetic: at $K{=}12$ the finest positive realized covered-gold-error bucket ($1/12 \approx .083$) exceeds the 5% budget, so only zero-drop items realize $\leq 5\%$ risk, while at $K{=}50$ the buckets $1/50{=}.02$ and $2/50{=}.04$ become realizable (populating them requires fresh $K{=}50$ draws, since the plug-in estimate is itself $1/12$-quantized). Finer K also pins the served mean: the bootstrap CI of the top-half served covered-gold-error narrows $\approx 8\times$ at $K{=}50$ (ts v3 $[.708, .757] \rightarrow [.730, .735]$; wt v2 $[.670, .715] \rightarrow [.689, .695]$) with the mean unchanged, so the flag-gate serving deltas are limited by K-concentratable noise, not by a systematic K effect.

D MENU GEOMETRY AND THE LIMITS OF ERROR DETECTION

This appendix varies which candidates are shown or withheld, then separates the errors the off-menu flag detects from those it misses. All calculations use the recorded per-item score vectors. Score demotion is a counterfactual diagnostic, not a deployable correction rule. "Gold rank" and correctness-based partitions use evaluation labels only. The figure at the end retains all 21 generator–suite cases, including negative and uncertain associations.

Terminology. A generator–suite *case* is called a *leg*; the nine *settled* legs are the four CSQA-941 legs, the four ts v3 legs, and the wt v2 host leg. The *wrong union* consists of items answered incorrectly by at least one of the four generators. The *core* contains items wrong on all four; the *peripheral* class contains items wrong on exactly one. An item's depth is its mean leader-minus-gold score gap over its wrong cases. "Heard" denotes the union of flag detection and the recorded score-demotion response at the stated δ; "unheard" denotes its complement. These are diagnostic partitions, not semantic judgments or achieved answer repairs.

D.1 HOW DOES SHOWING FEWER OR MORE CANDIDATES CHANGE THE THREAT?

Table 7 serves each family's full $M{=}5$ established menu; here (Table 12) the provider chooses a *sub-menu* $\{\text{gold}\} \cup S$ with S a size-k subset of the four native distractors, all $\binom{4}{k}$ sub-menus enumerated per item ($k{=}1, 2, 3$; no model calls — the established-menu LL analysis already stores, per item, the summed logliks of all five menu options and of the extended off-menu candidates on the same 941 items (Appendix E.8's basis match to the masked pass; source results/asm_csqa941_menu_subset_ablation.json). Confidence intervals are paired item-bootstraps ($B{=}2000$ shared resamples, seed 20260903); each item is resampled together with all its sub-menus. An excluded candidate *outranks* the presented leader when its conditional loglik exceeds the presented maximum; the excluded set is the hidden native options plus the extended off-menu candidates.

Synthetic suites. The enumeration ports unchanged to the synthetic masked-mass suites — ts v3 for all four generators and WikiText-103 v2 for g1, the suites whose per-item extended-set masked conditionals are stored (source results/asm_synthetic_menu_subset_ablation.json). Under strict outrank the uniform order is again monotone in M on all 5 legs; the strict $M{=}5$ threat levels are .297/.372/.281/.305/.367 and the $M{=}2{-}M{=}5$ deltas are $+3.1[2.4, 3.9]$ / $+3.2[2.5, 4.0]$ / $+2.4[1.8, 3.1]$ / $+3.2[2.5, 4.0]$ pp on ts v3 (g1 / g2 / SMOLLM2 / Qwen3-4B) and $+1.0[0.7, 1.5]$ pp on wt v2 (same paired bootstrap). The published $\tau{\geq}1$ flags count exact-tie items (49 across the five legs); the rates here use the strict convention, which is what the sub-menu comparisons need.

Grown menus. The last geometry moves in the opposite direction — add k off-menu candidates to the five listed options. All numbers below are exact rank combinatorics on the per-item 21-vectors (no enumeration needed; the uniform-law subset counts are verified against direct enumeration at $k{=}1, 2$), still on the same legs.

In summary, shrinking the menu cannot erase the threat and growing it uniformly only discounts it; full admission removes the threat but also the answer accuracy; selective serving remains the only option that both keeps the answer and measures the hostage share.

family	P(excluded competitor outranks presented leader) @M=2	@M=3	@M=4	@M=5	adversarial
g_1 Qwen2.5-1.5B	.768	.752	.737	.721	.784
g_2 Qwen3-0.6B	.958	.954	.951	.947	.962
SmolLM2-1.7B	.922	.907	.892	.877	.937
Qwen3-4B	.804	.794	.784	.774	.814

Table 12: $M \neq 5$ sub-menu cells on the established CSQA-941 menu (all four generators; LL basis, 941 items each). The uniform-over-sub-menus outrank threat changes little with M: it swings by at most 4.8 pp within each family across $M \in \{2, 3, 4, 5\}$ — paired item-bootstrap 95% CIs on the M=2 − M=5 delta: $+4.8[3.7, 6.1]$ / $+1.1[0.6, 1.8]$ / $+4.5[3.4, 5.7]$ / $+3.0[2.1, 4.0]$ pp — and stays monotone in M for every family (shrinking the menu cannot erase the threat). Hiding native options adds its own channel (M=2 rates .255/.407/.390/.260 for the four rows), yet beyond the free-candidate channel the increment never exceeds .009: the outranker is almost always an extended off-menu candidate. The adversarial column is an order-statistic identity — the provider needs to withhold only one target, so its rate is M-constant and equals P(extended-set argmax $\neq$ gold); the equality holds per item on all 11,564 items of the nine measured legs, cross-checked against the grown-menu serve-all column and the synthetic sub-menu cells (`results/asm_adversarial_orderstat_check.json`). On the only view-sensitive family the first-token view moves the flag column by +7.0 pp and every sub-menu column by at most 5.9 pp, with uniform monotonicity and every CI's exclusion of zero unchanged (`results/asm_csqa941_msub_viewcheck.json`).

suite	family	flag @ M=5	outrank threat @ k=8 uniform	outrank threat @ k=8 adversarial	served argmax acc uniform	served argmax acc by-order
CSQA-941	g_1 Qwen2.5-1.5B	.721	.360	.721	.277	.216
	g_2 Qwen3-0.6B	.947	.473	.947	.054	.038
	SmolLM2-1.7B	.877	.438	.877	.091	.063
	Qwen3-4B	.774	.387	.774	.233	.186
ts v3	g_1 Qwen2.5-1.5B	.297	.148	.297	.739	.661
	g_2 Qwen3-0.6B	.372	.186	.372	.663	.585
	SmolLM2-1.7B	.281	.140	.281	.762	.687
	Qwen3-4B	.305	.152	.305	.725	.652
wt v2	g_1 Qwen2.5-1.5B	.367	.182	.366	.703	.619

Table 13: Grown-menu cells on the nine legs (source `results/asm_grown_menu_cells.json`). Serving k=8 off-menu candidates uniformly does not remove the threat: it decays by exactly the tie-class combinatorics (w^* tie members must all be admitted; law verified by enumeration), to roughly half the M=5 rate on every row. An adversary who keeps w^* out loses nothing through k=8 on any leg (adversarial column = flag column, up to pool-exhaustion rounding on legs with <8-candidate pools). Presenting the full pool by score order eliminates the threat (no excluded candidate remains to outrank the leader) at the cost of pool-argmax accuracy: the by-order column is 25–47pp below the menu argmax on all nine legs. Paired item bootstrap (B=2000) 95% CIs are strictly negative per leg: $-44.4[-47.7, -41.3]$ / $-41.9[-45.2, -39.0]$ / $-41.8[-45.0, -38.8]$ / $-46.8[-49.8, -43.7]$ / $-26.4[-28.7, -24.1]$ / $-30.9[-33.3, -28.6]$ / $-24.6[-26.8, -22.4]$ / $-25.0[-27.2, -22.9]$ / $-33.4[-35.6, -31.3]$. CSQA-941 legs use the length-normalized view, in line with the sub-menu cells above; on the only view-sensitive family the first-token basis moves the flag column by +7.0pp (.877 → .947) and the by-order drop from 41.8pp to 47.8pp, leaving the qualitative cells — uniform halving, adversarial intactness, strict-negative drops — unchanged (`results/asm_grown_menu_cells_smol_viewcheck.json`).

D.2 What changes when the gold answer is withheld?

Unserved-gold menus. One further geometry is measured here: the provider omits the gold and prints k of the four distractors. Since every artifact stores the gold's own masked score inside its 21-vector, these menus are again a pure re-derivation on the same protocol (4×941 CSQA-941 items and the five synthetic legs; strict outrank; anchors cross-checked on g_1).

suite	family	flag (served $M{=}5$)	outrank threat (uniform) $M{=}2$	$M{=}3$	$M{=}4$	gold leads excluded	rescue (gate)
CSQA-941	g_1 Qwen2.5-1.5B	.721	.984	.968	.952	.856	.219
	g_2 Qwen3-0.6B	.947	.996	.993	.989	.736	.038
	SMOLLM2-1.7B	.877	.985	.970	.955	.760	.064
	Qwen3-4B	.774	.990	.980	.970	.851	.188
ts v3	g_1 Qwen2.5-1.5B	.297	.989	.979	.968	.974	.668
	g_2 Qwen3-0.6B	.372	.989	.978	.968	.964	.591
	SMOLLM2-1.7B	.281	.992	.984	.976	.978	.693
	Qwen3-4B	.305	.989	.978	.968	.966	.659
wt v2	g_1 Qwen2.5-1.5B	.367	.996	.993	.989	.985	.622

Table 14: Unserved-gold sub-menu cells. The provider omits the gold and presents k of the four distractors; "gold leads excluded" is the share of (item, sub-menu) pairs whose strongest excluded candidate is the gold ("rescuing" makes it conditional on a flag: $P(\text{gold-led} \mid \text{outrank})$, computed cell-wise as $\sum(\text{led} \wedge \text{outrank}) / \sum(\text{outrank})$ over (item, sub-menu) pairs; the ratio of marginal shares would over-count, because gold-led does not imply outrank on the same pair: by up to 0.7 pp on the displayed $M{=}2$ column and 2.3 pp across all cells, and on ARC that ratio exceeds 1.0, reaching 1.88). Threat columns are near-saturated: at $M{=}2$ a uniform provider is outranked on .984–.996 of pairs on every leg, and an adversarial provider reaches 1.0 on all nine (degenerate). The gold-led column splits by suite: on the established menu the strongest excluded candidate is typically *not* the gold (.218–.188), so the gate's deferral restores the gold in only .219–.188 of flags; on the synthetic legs the gold dominates the excluded set (.666–.621) and rescue reaches .668–.622. A protective provider's gold-threat at $M{=}2$ equals the rate at which the gold strictly leads all distractors — the served accuracy up to five exact-tie items across the nine legs (verified per item). On the only view-sensitive family the first-token view moves no column of this table by more than 3.4 pp (`results/asm_unserved_gold_smol_viewcheck.json`). Source: `results/asm_unserved_gold_cells.json`.

The unserved-gold geometry therefore extends, not overturns, the picture: gold-omission is uniformly hostile — selective service keeps flagging nearly every menu on all nine legs — but what the flag rescues depends on the suite: on the established menu it mostly restores another candidate (the under-inclusion signature persists even when the gold is withheld), while on the synthetic suites it restores the omitted gold in a majority of flags.

Pool size (K-truncations). The served flag at $M{=}5$ compares only the single strongest excluded candidate, which every top-K truncation with $K{\geq}1$ preserves, so the flag column of every table above is pool-size invariant; and the grown-menu uniform threat at fixed serve count decays combinatorially in pool size (Table 13). This analysis adds the structural measurement behind both: where the gold itself would rank among the off candidates (earliest rank; ties not counted). On the established menu the gold is the pool's best candidate in only 3.8–22.6pp of items, inside the top four in 9.1–45.9pp, and inside the full 16-candidate pool in 23.8–64.1pp (median rank 6–17); on the synthetic suites it is the best off candidate in 61.1–70.9pp (median rank 1 on all five legs) and inside the top four in 85.9–93.1pp. The under-inclusion signature is thus a pool-level fact, not only a served-menu one: the established menu's off pool is populated by distractors stronger than the gold, the synthetic pools by weaker ones. Pools truncated from samplers or menus other than the nine measured legs are not measured (source `results/asm_gold_offrank_profile.json`).

Join of the two surfaces. The gold-led channel of Table 14 is the conjunction of two dominance conditions: the gold tops its own generator's off pool (the rank-one cell of the profile above) *and* dominates the presentation's excluded natives. Recomputed cell-for-cell against the published unserved-gold cells (all 108 uniform cells matched at the stored rounding) and cross-checked against the profile's rank-one cells (all nine legs), the join shows that the rescue split *is* the off-rank split passed through a near-transparent filter: conditional on topping the off pool, the gold also dominates the three-native exclusion set of the $k{=}1$ presentations on 96.1–100.0pp of the established menu's items and 96.7–99.1pp of the synthetic ones (per-leg paired 95% bootstrap half-widths $\leq$2.3 pp and $\leq$1.0 pp, $B{=}2000$). Two statements are exact, not statistical: every gold-led cell on the full (item, presentation) grid is a pool-top event (zero violations on 11,564 items), and the gold-led share is nondecreasing in the presented count k on every item, a theorem of the shrinking-exclusion geome-

try (source `results/asm_rescue_offrank_join.json`). On the only view-sensitive family the first-token view shifts no cumulative profile cell by more than 3.4 pp (rank-one status changes on 24 of the 941 items; the median rank is unchanged) and moves the join conditional from 98.8pp to 100.0pp (+1.2 pp); the two exact statements are verified within each view (`results/asm_gold_offrank_viewcheck.json`).

Join with the served menu. The hostage mass sits on the correct, not the wrong: per item, correctness (served 5-option argmax = gold) cross-classified with pool-top status shows that on the established menu 67.3–91.6pp of correctly-served mass is pool-hostage (per-leg paired 95% bootstrap half-widths $\leq$3.7 pp established, $\leq$2.6 pp synthetic, B=2000; an off candidate strictly outranks the gold *while* the menu answered right), against 26.4–35.0pp on the synthetic suites. The reverse conditional is suite-uniform: pool-top items are served correctly at 95.3–100.0pp on all nine legs (half-widths $\leq$3.0 pp) and the native-loss-but-pool-top quadrant never exceeds 3.0pp — under-inclusion attacks exactly the population the menu gets right, and the suites differ only in how large that population is (consistency check: served flag/accuracy recomputed against the unserved-gold anchors, pool-top rates against the K-profile cells, 3×9 matches; source `results/asm_served_offrank_join.json`).

D.3 Which menu relationships transfer across benchmarks?

OpenBookQA. On the second established suite (Appendix F.5) the direction survives but the certificate degrades: the hostage share of correctly-served mass runs 43.1–75.9pp (CI half-widths $\leq$5.3 pp), always positive but below the CSQA range on three of four families, and pool-top certifies correctness only at 50.0–71.8pp — essentially chance for the two weakest families — while the native-loss-but-pool-top quadrant populates to 4.0–20.5pp (vs. $\leq$3.0 on CSQA). Read as a boundary of the mechanism, not a contradiction: in the deep under-inclusion regime pool and menu disagree independently, so the certification layer is suite-dependent even as the hostage direction is not (accuracy checked against the stakes cells of Appendix F.5 plus the stored fields of the base analyses, 2×4 matches; source `results/asm_served_offrank_join_obqa.json`).

The unserved-gold geometry replicates item-exactly on the same four OBQA legs: every gold-led cell is a pool-top event, the gold-led share is nondecreasing in k per item, the protective k=1 gold-threat equals strict-lead (zero ties on 4,000 items, vs five on the CSQA cluster's 11,564), and its uniform rescue at M=2 runs 10.7–23.3pp. What changes is only the weight of the join's second factor: conditional on topping the off pool, the gold dominates the two-native exclusion set of the k=1 presentations on 58.6–77.5pp of items, against 96.1–100.0pp on CSQA — in the deep regime the native filter, not the pool, is what under-inclusion starves (checked against the served anchors per item and the join analysis of Appendix D.2; source `results/asm_unserved_gold_obqa.json`).

Pool depth completes the cross-benchmark map (OpenBookQA, ARC). The earliest-rank K-profile completes on the second and third suites (all four families each, ll basis; consistency checked item-exactly: the profile's pool-top incidence reproduces the top-one rate of both join analyses). The gold's depth in its own off pool is a suite property with a clean ordering. Pool-top incidence runs 3.8–22.6 pp on the established menu and 61.1–70.9 pp on the synthetic legs; the second and third suites climb out of the established depth, in order: OpenBookQA at 14.2–41.0 pp (mean rank 6.3–8.7, median 2–7) and ARC at 53.8–82.7 pp (mean rank 2.2–3.2, median 1) — the leg range overlapping the synthetic profile, as that suite's rising rescue share in the unserved grid above already hinted. The certification layer tracks the depth ordering suite-wise: pool-top certifies correctness at 95.3–100.0 pp on the settled legs, degrading to 50.0–71.8 pp on OpenBookQA and 40.4–58.0 pp on ARC — where the gold usually already tops its own pool, the pool-top event carries little information about correctness. As a leg-level co-movement the relationship is descriptive only: pooled over the 17 legs the Spearman correlation of pool-top incidence with the certification rate is -0.39 ($[-0.81, 0.20]$, straddling zero) because suite membership dominates the units; within every suite the direction is negative (-1.00 OpenBookQA, -0.80 ARC, -0.22 on the nine settled legs, whose certification layer is near-saturated and leaves little room to co-move; source `results/asm_gold_offrank_profile_m4.json`).

The gate's precision–recall picture against the under-inclusion target. One item-exact identity closes the precision side: the served flag compares the strongest excluded candidate with the served leader, whose score weakly exceeds the gold's, so flag $\Rightarrow$ rank $\geq$ 2 holds on every one of the

25488 measured items across all five suites (zero violations) — the off-menu flag is a *perfect-precision* under-inclusion detector; it never fires on a pool-top item. Recall, however, is suite-ordered and depth-structured. The share of under-inclusion escaping the gate runs 1.6–5.7 pp on the synthetic legs and 1.5–6.9 pp on the established menu, but rises to 12.0–28.4 pp on OpenBookQA and 27.9–46.6 pp on ARC; the riddle legs slot into the same ordering at 6.2–31.8 pp, overlapping OpenBookQA's band while carrying the corpus's deepest under-inclusion pools (74.7–90.6 pp of all items). The escape is concentrated at shallow depth, by construction: at rank 2 the single superior off candidate hides whenever a beaten native still scores above it, and recall indeed climbs with depth exactly where the native mask matters — on ARC g_2 from 35.9 pp (rank-2 band) to 47.1 pp (rank-$\leq$ 8 band) and on OpenBookQA g_1 from 72.8 to 85.7 pp, while on the established menu and the synthetic legs recall sits at .90–1.00 pp essentially flat across bands. Equivalently, the flagged populations live at different depths: 16.1–40.6 pp of ARC flags already sit at rank 2 against only 2.5–12.2 pp on the established menu. Together, pool depth (the preceding paragraph) measures how much under-inclusion there is, this analysis measures what share of it the gate sees, and the certification layer measures how informative a raised flag is.

D.4 Why can a wrong menu leader hide an omission?

Score margins. A second item-exact identity follows from the same construction: on correct items with rank $\geq$2 the leader is the gold itself, so correct $\wedge$ rank$\geq$2 $\Rightarrow$ flag on all 25488 items — every escaped item is wrong, and the gate's decision there is one signed scalar, $m = \max_j \text{score}(\text{off}_j) - \text{score}(\text{leader})$. On the 7464 wrong-but-under-included items across the five suites the two sides of that scalar are suite-ordered *in opposite* directions: median catch clarity thins from 2.8–6.5 nats on the settled legs to 1.1–1.9 nats on ARC — with the riddle legs the clearest of all at 2.0–3.5 nats, matching their menu-depth measured below — while the median miss overhang (mask thickness) holds at 0.6–1.4 nats on the settled legs and 0.5–1.1 on OpenBookQA, thickening to 1.5–1.6 nats on ARC's three small families. It is not a razor's-edge phenomenon: mass within $\pm$0.25 nats of the boundary is only 2.5–18.9 pp of the maskable set on every leg, and the rank-2 escape identified above is by construction the presence of a beaten native above the single superior off candidate. The resulting design constraint for the deferral layer is quantitative: to surface ARC's hidden half of under-inclusion a gate would have to demote wrong-but-strong natives by well over a nat of score — recovering recall is not a recalibration of the flag but a re-scoring of the native field.

The demotion frontier measures the trade, not just the point. The mask-thickness point estimate of the preceding paragraph is extended here to a full counterfactual surface: demoting every native by δ nats moves the gate to $\text{flag}_i(\delta) := \text{offmax}_i > \text{leader}_i - \delta$, evaluated by exact arithmetic on the stored per-item scores against its coverage cost $\text{F}(\delta)$. At δ=1.5 — the median thickness above — the volume confirmation is exact: ARC's small families recover 49.6 pp of the previously hidden mass on g_2 and reach recall 76.5–84.2 pp across the three (OpenBookQA g_2: 91.7 pp) while the g_2 flag rate moves only 10.3 $\rightarrow$ 22.8 pp. The marginal exchange rate, however, is suite- and family-shaped: on ARC g_2 a +23.1 pp recall gain costs +12.5 pp of menus flagged, while on CSQA-941 g_1 the same demotion buys recall up to only 98.1 pp at a 72.0 $\rightarrow$ 89.6 pp flag rate — over three flagged menus per recovered point of recall — and on ARC g_1 recall saturates (97.5 pp) while the flag rate itself doubles (33.3 $\rightarrow$ 68.5 pp): the thin mask of the strongest family makes the same demotion costly in coverage rather than inexpensive. The riddle legs, carrying the corpus's deepest hidden pools, recover the most at the same line — 43.6–85.5 pp of the previously hidden mass for flag moves of 12.1–21.5 pp. (Basis artifact, sharpened by a string-level audit: the few misses at exact zero margin on the masked legs — up to 22.2 pp of each — are *not* duplicate strings (zero duplicate slots in both string-carrying pools) but 6 morphological twin pairs across the two legs whose shared leading subword aliases them to bit-equal stored first-token scores (bird/birdie, fluttered/fluttering, growled/growling): the masked basis cannot separate a same-stem competitor from the menu string, the item-level witness of the first-token view's measured sensitivity; source `results/asm_offmenu_dup_audit.json`. The ll legs tie only on a measure-zero set.) Consistency checks: δ=0 cells reproduce the flag rate and recall of the preceding analysis item-exactly on all 21 legs, the quartile-integration identities hold against the stored mask thickness, and all three curves are monotone by assertion (source `results/asm_flag_demote_surface.json`).

First-token aliasing, measured globally. The twin-pair audit above isolated the items where aliasing reached the gate; the same bit-equality criterion measures the view's total coarseness with the ll legs as coincidence floor. On the synthetic masked pools, 25.3–67.3 pp of *all* items carry at least one

score-equivalence class spanning menu and off candidates (510–1371 classes per leg, up to 20 tied candidates in one class) — the short-story vocabularies merge into same-stem clusters whose size tracks the tokenizer, not the suite (25 pp on g_2 against 67 pp on the 4B leg of the identical pool). The ll legs run the identical procedure and find 1 collision in 13604 items: the masked excess over that floor is the structural cost of reading first tokens only. Two containment facts keep the reading above intact: the gate-relevant intersection — aliasing that involves the served leader's own score — covers only 0.7–1.4 pp of items (78 across the five masked legs), and all six exact-zero-margin misses of the demotion surface sit inside it (asserted per leg where strings are stored). The view-basis sensitivity of 3.4 pp aggregate on the one view-sensitive family therefore has a measured microscopic substrate: frequent local merges, rare gate-implicating ones.

D.5 Which errors are reachable, and what does deferral cost?

Read against the correctness target, the served error mass splits by construction into a contained channel (wrong and rank $\geq$2: an excluded candidate out-scores the gold, so an off-pool gate can see the item) and an unreachable floor (wrong and rank 1: the model is wrong with no better offer anywhere outside the menu, which no amount of off-pool scrutiny — or demotion — could have separated). The floor is suite-ordered and decisive: 0.0–3.1 pp of errors on the established menu (exactly 0 on g_2: every error there has a superior excluded candidate), 6.9–32.0 pp on OpenBookQA, and 42.6–76.9 pp on ARC, and only 1.1–6.0 pp on the riddle legs — the corpus's hardest benchmark carries nearly no unreachable error at all: almost every riddle mistake has a superior excluded offer, so the gate's channel is wide open there. On ARC's small families, by contrast, *no* off-pool design can contain more than about a quarter of the error mass, whatever the threshold: demotion to δ=3 still reads only 10.9 $\rightarrow$ 44.0 pp of errors on g_2, against 81.2 $\rightarrow$ 99.4 pp on the established g_1 at δ=0 $\rightarrow$ 3. Composition agrees: on the first benchmark the flag concentrates error (clean service after deferral lands 77.2 pp correct against the base 66.0 pp), while on ARC the deferred and served populations read nearly identically (46.8 vs 46.9 pp) — the same gate is a wrongness concentrator on the deep-gold menu and a near-neutral repartition where the gold is shallow. Where the three preceding analyses measured the detector, this one measures its ceiling: demotion can open only the contained channel, and the unreachable floor needs a different surface entirely — answer-level repair, not off-pool scrutiny.

The marginal spectrum: what the coverage cost is made of. Each unit of demotion was evaluated above by recall and flag inflation; the identity chain splits every marginal flag flag(δ) $\setminus$ flag(0) into exactly three classes — rescued contained errors (wrong, rank $\geq$2: the signal), unreachable wrongs (rank 1: pure cost), and deferred good service (correct, rank 1: the silent cost). At δ=1.5 the signal share re-orders the suites for the last time — and the head of the spectrum is new: the riddle legs run 38.2–68.5 pp (68.5 on g_2, and 77.8 pp already at the first quarter-nat step on SmolLM2), the corpus's richest marginal trades, on its deepest pools. OpenBookQA follows at 44.2–54.8 pp as the stable middle instrument, passing ARC at 32.9–37.1 pp, and the synthetic pools collapse to 2.3–7.4 pp — where recall was already saturated, so the same demotion drags in rank-1 near-boundary mass at up to 42.4:1 junk per rescued error. The ARC g_1 coverage catastrophe decomposes the same way: of its 514 marginal flags, only a third are rescued errors while 107 unreachable wrongs and 236 correct items ride along. Deferral by demotion is therefore a middle-suite instrument on this corpus: worth buying where hidden mass exists, pure waste where the gate was already saturated. At the δ=3 endpoint riddle still heads the spectrum (32.1–67.4 pp) and the native bands hold below it — OpenBookQA 33.9–44.0 pp, then ARC 21.1–24.2 pp, with the synthetic pools pinned under five points — but the legs begin to cross inside the native suites (CSQA's SmolLM2 leg at 45.8 pp overtakes OpenBookQA's at 33.9 pp; riddle's g_1 at 32.1 dips under that same band), and the synthetic tail is by then almost entirely drag: up to 51.2:1 junk, 90.8–96.5 pp of its marginal flags already-correct service riding along. Bootstrapped item-level (B=2000), the head separation survives at interval level, if only just: riddle's two head legs run $[62.1, 74.4]$ pp (g_2) and $[61.2, 74.1]$ pp (SmolLM2) against OpenBookQA's head leg $[48.6, 61.0]$ pp, and the synthetic collapse is decisive (every synthetic interval ends under 11.3 pp); the δ=3 interior crossings are within noise (CSQA's SmolLM2 leg $[36.3, 55.8]$ vs OpenBookQA's $[29.6, 38.4]$ pp overlap).

D.6 How large are the score gaps in missed errors?

Appendix D.5 assigned the unreachable floor to a different surface (answer-level repair); measuring the floor shows that most of it does not need one. On every item of that floor a single scalar exists, the leader's menu-internal overhang of the gold, gap $=$ score(leader) $-$ score(gold) — and the menu-side twin of the off-side mask thickness turns out thin: medians run 0.40–0.86 nats on ARC (0.40 on g_1, whose floor contains 81.5 pp within one nat), and 83.3–99.1 pp of ARC's unreachable errors sit inside two nats (70.9–92.5 OpenBookQA, 68.9–92.5 synthetic). The riddle legs hold the only caveat: their floors are few (6–39 items each) but the thickest in the corpus — medians 0.64–1.94 nats, with by leg 52.9–100.0 pp inside two nats — the one suite whose rare unreachable errors can be genuinely deep. A string-level audit of all 96 of them confirms the depth is genuine: only 3 carry a leader sharing a twin prefix (Appendix D.4) with the gold, and those are template-shared distractor *sentences*, not word-morphology twins; 31 are thick (≥ 2 nats) with only 2 twins among them — confident semantic wrongness (straw options adversarial by design; the Qwen3-4B leg floors "bird" under "weeping willow" by 8.08 nats, 47.1 pp of its floor thick). The big native floors pass the same audit with a sharper lesson: 13.7–19.1 pp of ARC floor items share at least half their tokens with the gold, and exactly that near slice carries the thinnest margins (medians 0.20–0.47 nats against 0.46–0.96 for the distinct majority; 2.1 pp of g_2's floor are exact token-multiset inversions — ordering questions whose semantics live in the order); OpenBookQA's floors run 5.0–10.0 pp token-near with no exact inversions (medians 0.24–0.87 vs 0.52–1.32). The thin native floor is thus structural near-tie — contrast pairs and inversions built into the menus — not label noise. The two remaining tails close under the same audit: CSQA's floor is tiny by alignment (19 items across the four legs, mixture of menu-specific surface pairs and genuinely adversarial ones), and the synthetic floors are pool-substitution rather than crosstalk — of 167 floor items across the five synthetic legs not one is token-near or a token-inversion, and the only shared-prefix twins are the "out"/"outside" pair, seen again in both directions (5 items) — the synthetic thin floors are pool-internal near-misses, not string artifacts. Equivalently, the degenerate demotion counterfactual — demoting every native by δ, which never disturbs a correct item since the leader there is the gold — shows the floor to be a bounded margin phenomenon: accuracy at $\delta{=}2$ would stand at 79.9 pp on ARC g_2 against 35.9 pp unswept, a $+44$ pp ascent carried almost entirely by items the off-pool gate is blind to. (This is a depth profile, not a proposed policy — a uniform constant prior for the gold is statistically meaningless as a serving rule; the point is that a near-tie-aware, calibration-aware *menu-internal* correction is in range of the measured margins, while the off-pool gate analyzed above cannot reach them at all.) Together with the off side, the geometry is thus measured on every suite: pool depth measures under-inclusion, mask thickness measures the gate's view of it, and the floor gap measures the part it cannot see.

The contained class on the same ruler: what the gate uniquely owns. The question left open by the last two analyses is sharp: if the floor and most margins are thin, does a menu-internal correction make the off-pool gate redundant? Measured on the same $\delta{=}2$ ruler, the answer splits the contained class itself. Its menu-side depth is shallow but not floor-shallow on the four non-riddle suites: medians run 1.11–1.54 nats on the synthetic pools, 1.16–2.60 on ARC, 1.53–2.61 on OpenBookQA and 1.60–2.91 on CSQA, with by leg 16.1–72.8 pp of contained errors inside two nats — calibrator-substitutable in principle. The riddle legs anchor the other end of the map: contained medians 3.06–5.88 nats (the corpus's menu-deepest regime), by leg only 16.1–34.1 pp inside two — and recall from the partition that this is also the suite whose errors are almost all contained (floor 1.1–6.0 pp). The gate's exclusive territory is the complementary menu-deep cell: 11.0–57.1 pp of contained errors are deep *and* already caught — deep violations the gate prevents without touching anything else — headed by riddle (42.4–57.1 pp) and CSQA, and smallest on ARC where error is uniformly thick. The residual hard core, deep *and* escaped, runs 16.2–39.5 pp on ARC and 8.8–38.7 pp on the riddle legs — the cell no within-two-nat instrument reaches from either side. The complete menu-side counterfactual moves accordingly: demoting every native by $\delta{=}2$ flips not just the floor but the thin contained items too, lifting ARC g_2 to 87.4 pp (against 79.9 pp floor-only) — most native error on the four non-riddle suites is shallow on the menu side, so the gate's measured claim narrows: not unique repair of error generally, but a deployable item-local rule whose exclusive cell is menu-deep wrongness — precisely riddle's home ground — while the near-tie share remains a calibration phenomenon (consistency check: $n{=}7464$ contained items reproduce the error-partition counts item-exactly on all 21 legs, floor-cured mass and escape shares reproduce the floor-gap and mask-margin analyses; source `results/asm_menu_gap_contained.json`).

Uncertainty. An item-bootstrap layer (B=2000, the analysis seed, percentile intervals) sits under the five headline shares of this arc on all 21 legs, and the suite separations above survive at interval level: the thinnest contained share on riddle (16.1 pp on Qwen3-4B, interval 13.1–19.3) stays clear of the non-riddle minimizer (CSQA Qwen3-4B 35.2 pp at 29.9–40.2); the gate-exclusive cell on riddle g_1 (57.1 pp, 52.9–61.1) sits nowhere near ARC's (11.0 pp, 8.1–13.9); and the corpus's worst escape rate (ARC g_2, 56.2 pp) is bracketed at 49.8–62.7. The widest interval is exactly where the layer says it should be: the few-item riddle floors — Qwen3-4B's two-nat share brackets 35.3–70.6 pp at n=34 — so the thick-tail claim there rests on the deterministic string audit (every item inspected), not on frequency.

D.7 Which errors persist across generators?

The preceding comparisons pool the four generators of each suite; the backbone analysis asks which errors persist across generators by aligning the four generators item-by-item and splitting each wrong set into the core (wrong on all four) and the peripheral (wrong on exactly one). The backbone share of the wrong union is suite-ordered and instructive: 12.1 pp on the synthetic pools (the smallest — most of synthetic error is generator-level noise, itself a cause of the junk-dominated marginal spectrum), 23.6 pp on the established menu, 41.9 pp on both ARC and the riddle legs, and 51.0 pp on OpenBookQA (its pairwise conditionals run .75–.90). The classes of the preceding analyses lock onto their expected sides: the gate-exclusive cell is backbone-concentrated wherever it exists — 52.7 vs 20.8 pp core/peripheral on the riddle legs, 53.3 vs 23.0 on the established menu, 27.4 vs 15.3 on OpenBookQA — and flat where it does not (ARC: 6.4 vs 6.5 pp, the uniformly thin suite), while thin error rides the peripheral tail (15.6 vs 66.2 pp riddle, 91.0 pp of synthetic peripheral pairs). Menu-deep wrongness is therefore a property of the *item*, witnessed across generators, and the per-generator spread of the thin shares reported above is the same fact read from the other end: near-tie error is largely a generator-dependent outcome. Intervals computed the same way (item-level bootstrap, B=2000) confirm all three patterns: the share ordering is two clean separations with a point-identical tie — ARC $[39.2, 45.0]$ and riddle $[38.6, 45.1]$ pp coincide, synthetic and the established menu end at 16.1 and 26.9 pp below their lower endpoints, and OpenBookQA starts at 47.6 pp above their upper endpoints — the deep-caught lock gap excludes zero on the four suites that have the cell (lower endpoints $+5.0$ to $+25.0$ pp) and is a tight band around zero on ARC ($[-4.1, +3.5]$ pp), and the thin asymmetry is positive on all five (lower endpoints from $+15.5$ pp up). The generator agreement behind the item-property reading is assessed identically: the weakest pairwise conditional on OpenBookQA runs $[.71, .78]$ and the strongest on the riddle legs $[.90, .94]$, while the synthetic pairs stay decisively below the native suites ($[.26, .40]$–$[.61, .75]$) — agreement strength is suite-ordered in the same direction as the backbone share itself. The mechanism is dosage-driven on every family: taking item depth as the mean menu gap over the item's wrong legs, the backbone-core rate climbs across depth quartiles from 2.9–21.9 pp in the shallowest quartile to 41.1–74.8 pp in the deepest on the native pools, and from 1.5 to 30.9 pp on the synthetic pools, with every deepest-to-shallowest contrast strictly positive at interval level (lower endpoints $+17.4$ to $+61.9$ pp, B=2000 item bootstrap, quartile edges recomputed per replicate). The synthetic flatness is therefore positional rather than constitutional — contrary to the expectation that the synthetic curve would be flat, it climbs: synthetic wrong mass sits at shallow depth (median 1.0 nats against 0.8 on ARC and 3.5 on the riddle legs), so only a small slice of it ever reaches the depths at which wrong items flip generators; on ARC the union is still shallower yet the mid-curve is steep, showing the share ordering is a curve-height-by-distribution interaction, not depth alone.

Conditioned on spread: the carrier is pooled intensity. The dose curve above could still ride on a cruder lever — the wrong-leg count, a coordinate degenerate against the core label. Measured cleanly inside the spread-conditioned cell (items wrong on at least two legs), the intensity coordinate still decides saturation: core rates run 8.4–37.5 pp in the shallowest of the within-strata depth quartiles to 55.0–80.6 pp in the deepest on the native pools (8.6 to 54.3 pp on the synthetic pools), every deepest-to-shallowest contrast positive at interval level (lower endpoints $+22.9$ to $+54.9$ pp, same bootstrap design). Pooled, not peak: among these items the mean wrong-leg gap outpredicts the single deepest gap decisively on all five suites (ΔBIC $+11.5$ to $+112.6$ in the mean's favor), and the peak gap adds back signal only on CSQA and the riddle legs. Spread itself is intensity-driven: depth medians rise strictly with the wrong-leg count on every suite (asserted per item in the analysis), e.g. $0.6 \to 2.7$ nats on the synthetic pools and $1.2 \to 5.1$ nats on the riddle legs.

D.8 Do the flag and score demotion detect the same errors?

Errors shared across generators. The two preceding analyses meet at the instrument question: when demotion buys back an escaped contained error, is it a backbone trap or a peripheral coin flip? On all five families the rescue rate *inside* the backbone exceeds the peripheral one at δ=1.5 — 7.2 vs 5.1 pp on the established menu, 21.6 vs 18.1 synthetic, 24.2 vs 16.4 OpenBookQA, 33.7 vs 24.1 ARC, 18.0 vs 15.9 riddle — and the rescued pool is core-enriched over the wrong baseline by 3.0–8.4 pp. Two complements: the *still-escaped* mass is likewise backbone-dominated (37.5–78.6 pp core — deeper traps wait farther out the same frontier, cf. the demotion surface), and on the synthetic pools most rescued mass is peripheral noise either way (consistent with their junk-dominated marginal spectrum). This adds a mechanistic clause to the instrument reading of this appendix: on the native suites deferral by demotion buys *traps* preferentially, on the synthetic pools it buys noise preferentially — the signal-share ordering of the marginal decomposition is the backbone difference read in the other currency. The point ordering is tested by item-level bootstrap within suite (B=2000): the δ=1.5 core-minus-peripheral gap excludes zero on OpenBookQA ($[+3.5, +12.0]$ pp) and ARC ($[+3.9, +15.2]$ pp), brushes it on CSQA and riddle ($[-0.4, +4.5]$, $[-1.1, +5.4]$ pp), and is uninformative on synthetic ($[-8.7, +16.9]$ pp); at the δ=3 endpoint the lower endpoints of OpenBookQA, ARC and riddle are all positive (smallest $+1.7$ pp), while CSQA and synthetic still straddle zero. The "buys traps preferentially" claim is therefore decisive exactly on the two thin-native families and directional elsewhere. The rescue surface itself is depth-graded: among rescue-eligible items the any-leg rescue rate climbs across depth quartiles on every suite at δ=1.5 (4.0–27.5 to 25.1–60.8 pp on the native pools, 9.8 to 41.2 pp on the synthetic pools; deepest-to-shallowest contrasts positive at interval level, lower endpoints $+11.1$ to $+29.3$ pp, B=2000 item bootstrap with quartile edges recomputed per replicate). And the trap preference is not explained away by depth: at fixed depth the core-minus-peripheral gap still excludes zero in the deep stratum on all four native suites (lower endpoints $+2.0$ to $+15.3$ pp) and in the shallow stratum on three of the four (CSQA brushes zero there; the synthetic cell is underpowered as in the backbone CIs). The pattern becomes stronger at the δ=3 endpoint: the core preference in the deep stratum excludes zero on all five suites there (lower endpoints $+0.3$ to $+16.3$ pp, the synthetic interval brushing but not crossing zero), the rescue-rate depth gradient steepening to $+18.9$–$+46.4$ pp lower bounds — deferred demotion buys deep traps, preferentially, along both axes at once. The two instruments do not meet: on the wrong union the off-pool flag fires on 42.7–99.4 pp of items with the reach suite-dependent (98.5–99.8 pp of CSQA wrongness flagged and 92.1–95.3 pp of riddle wrongness, against only 39.9–45.6 pp of ARC's, Wilson intervals off the same counts), yet flag-fire and demote-rescue anti-associate or stay independent — the odds-ratio intervals end below one on CSQA, OpenBookQA and riddle (upper endpoints 0.31, 0.57 and 0.42) and bracket one on ARC and the synthetic pools. Nor is the union flat: at δ=1.5 the two instruments together cover 99.9% of CSQA's wrong union and 98.2% of riddle's (Wilson interval 99.2–100.0%), 93.4% on OpenBookQA, but only 79.0% on the synthetic pools and 56.2% on ARC (53.3–59.0%) — the unheard cell is 42.1% of ARC's wrong union at the δ=3 endpoint while CSQA's closes structurally to zero, and the hardening moves no suite's union by more than 1.8 pp; the polarity is a hearing split, not a routing relabel.

Which errors do both diagnostics miss? The unheard cell is not the deep one the depth story would suggest: on ARC its wrongness sits shallower than the heard union's (median 0.41 nats; difference -0.87–-0.60, the interval excluding zero, OpenBookQA's unheard cell likewise -1.33–-0.89) with the exterior wall much farther (best off-menu margin median -5.6 nats, difference -2.7–-2.2), the gold already the best excluded continuation (exterior rank median 1) and slightly less backbone-core (share difference -0.14–-0.03) — shallow wrongness with no exterior ladder, which neither a ladder-driven instrument nor an off-pool gate can reach by construction; the two deep-anchored suites' tiny unheard cells lean deep instead (5.49 nats on CSQA's single item, 2.01 on riddle's sixteen), too few to estimate and recorded as such.

D.9 How does diagnostic coverage vary with error depth?

The split is depth-graded on the thin natives: quartile by quartile of within-suite error depth the heard share on ARC climbs 27.0% (22.2–32.5) to 82.8% (78.0–86.7) from the shallowest to the deepest quartile (Figure 5; $+55.8$ pp, monotone; OpenBookQA 83.1% to 97.5%, $+14.4$ pp; synthetic $+11.8$ pp), while CSQA stays at 100.0%–99.4% in every quartile and riddle's tilt is $+2.3$ pp

— the instruments hear precisely the deep, consequential wrongness there and fail to detect errors only in the shallow half.

Depth versus cross-generator agreement. The gradient belongs to the gap coordinate rather than to generator consensus: read against the backbone count the same endpoints on ARC are 49.7% (k=1, 42.9–56.6) to 61.2% (k=4, 56.8–65.5), a $+11.5$ pp tilt against the gap axis's $+55.8$, and the lack of detection is not a peripheral-tail effect — even fully-consensus ARC wrongness goes unheard at 38.8% (0.0% on CSQA, 0.5% on riddle, 5.9% on OpenBookQA), while the synthetic legs invert the k-ordering (78.9% to 72.7%, -6.2 pp). The contrast is estimated with intervals on the same items (item bootstrap B=2000): ARC's $+11.5$ pp move sits at $[+3.4, +19.8]$ — positive but nowhere near the depth axis's $[+48.4, +61.9]$, with its k=4 lack of detection at $[34.4, 43.4]\%$; riddle's $+4.6$ pp $[+1.3, +8.6]$ also excludes zero, while OpenBookQA $[-2.3, +8.2]$ and synthetic $[-24.3, +9.5]$ straddle it.

Conditioning on generator agreement. Nor is the steep gap-gradient a consensus-composition artifact (the outer quartiles *do* trade k-dominance — ARC's shallowest runs 38.6% k=1 against 11.9% k=4, its deepest 69.1% k=4): re-cutting the depth quartiles *within* every fixed-k stratum leaves each ARC stratum steep ($+54.0$/$+48.1$/$+62.9$/$+63.1$ pp for k=1–4, every interval excluding zero; standardized $+57.0$ pp $[+49.7, +64.8]$; even the fully-consensus stratum climbs $+63.1$ pp $[+52.0, +73.2]$), and OpenBookQA's move likewise survives standardization ($+15.0$ pp $[+9.3, +22.2]$). Composition control also unmasks the pooled-undecidable synthetic gradient ($+17.3$ pp $[+2.6, +32.2]$ within strata; the mix swing is widest there: 82.4% k=1 in the shallowest quartile against 23.5% in the deepest) and every conclusion here is unchanged under item-weighted pooling (synthetic $+15.1$ pp $[+1.7, +28.0]$), while the deep-anchored pair stays flat within strata (-0.4/$+0.2$ pp).

Window and depth-summary sensitivity. And the shallow lack of detection is not a window choice: tripling the rescue window moves ARC's three shallowest quartiles by 0.0 pp exactly while only the deep end gains ($+7.0$ pp, to 89.8%; OpenBookQA's tail $+2.0$ pp, synthetic $+2.9$, CSQA's cell closing to 100.0%) — wall placement, not tuning. Nor is the collapse load-bearing: replacing the mean gap over wrong legs by the max or the median leaves every suite in its named class — ARC $+54.4$/$+55.8$ pp (mean $+55.8$), OpenBookQA $+11.4$/$+15.4$ pp, the deep-anchored pair within -0.6..$+2.8$ pp — with item-depth ranks nearly unmoved (Spearman 0.991–1.000 vs. the mean).

Uncertainty and individual-generator patterns. The moves carry intervals, not just point estimates (item bootstrap B=2000, quartile edges re-cut per replicate): ARC $+55.8$ pp $[+48.4, +61.9]$ with a weakly monotone curve in every replicate (1.000), OpenBookQA $+14.4$ pp $[+9.0, +20.4]$ — both exclude zero. Riddle ($+2.3$ pp $[-0.5, +5.1]$) and the synthetic pool ($+11.8$ pp $[-1.5, +25.6]$) straddle it, and CSQA's near-zero move brushes it from below (-0.6: $[-1.7, +0.0]$) — the steep list is exactly {ARC, OpenBookQA}, the flat list {CSQA, riddle}, synthetic unresolved. Nor is the gradient one leg's pocket: leg-wise curves on ARC sit at $+55.2$..$+59.7$ pp, each weakly monotone, OpenBookQA at $+9.7$..$+14.5$ pp with every leg positive, while CSQA (-0.8..$+0.0$ pp) and riddle ($+0.0$..$+2.2$ pp) are flat on all four legs and the synthetic legs lean positive ($+5.4$..$+22.5$ pp, unsigned) — the gradient, not one generator's mask quirks.

Which diagnostic contributes to the depth pattern? The gradient's instrument mix rotates everywhere, not only where the total moves: ARC's climb splits across all channels (flag $+30.2$ pp, rescue $+48.8$, overlap $+23.2$ pp), OpenBookQA's is carried by the overlap ($+28.4$ pp) while the flag-only share thins, and even the saturated pools rotate their overlap cell — $7\rightarrow43$ of 175 items on CSQA ($+20.6$ pp), $35\rightarrow108$ of 218 on riddle ($+33.5$ pp) — with total coverage unmoved: the deep-anchored flatness is saturation, not instrument stasis. Those channel moves carry intervals (item bootstrap B=2000, quartile edges re-cut per replicate): the δ-band and the overlap rise on *every* suite (overlap $[+13.6, +28.3]$ CSQA, $[+24.8, +41.7]$ riddle, $[+7.1, +30.9]$ synthetic, $[+19.6, +35.8]$ OpenBookQA, $[+17.7, +28.4]$ ARC), while the flag-only climb excludes zero exactly on the gap-anchored pair (ARC $[+21.9, +37.5]$, OpenBookQA $[+4.5, +19.4]$) — the rescue side of the gradient is universal, the flag side is suite-typed.

Partition by the outside score margin. The mechanics can be expressed as a partition by the outside score margin: every zero-wall wrong item (an excluded continuation outscoring the leader on a wrong leg) is flagged on every suite, item-exactly (and the flag then hears another 3.7–10.1% of each union on top, firing on legs the model answers right); the δ-window spends the near band $(-\delta, 0]$, and the unheard cell is almost entirely gold-topped — $53/53$ on OpenBookQA's, $56/57$ on the synthetic pools', $495/499$ on ARC's and $11/16$ on riddle's unheard items already rank the gold first among excluded continuations, so no ladder exists for demotion to surface, most of them far-walled besides ($386/499$ of ARC's) (source `results/asm_wall_partition.json`; the near⊆flagged identity is asserted per suite). The flag is systematically quieter on the rescue-eligible items (P(flag | rescued) minus P(flag | non-rescued) runs -18.0–-7.4 pp on OpenBookQA and -12.5–-4.9 pp on riddle), and the rescue surface sits preferentially where the flag is silent on the deep native suites (P(rescued | flagged) minus P(rescued | flag-free) -32.3–-13.7 pp on OpenBookQA, -48.1–-21.2 pp on riddle; CSQA's flag-free cell is only four items, bracketed -90.1–$+11.7$ pp). The route split is a depth story: on the two suites with enough silent mass to estimate it, flag-free wrong items are systematically shallower (medians 0.59 vs 0.94 nats on ARC, 1.24 vs 1.71 nats on OpenBookQA; differences -0.53–-0.22 and -0.86–-0.04 pp, both excluding zero), riddle's and CSQA's silent cells lean deeper but are too few to estimate — the off-pool flag hears the deep wrongness demotion cannot reach, while the silent cell it misses is the shallow near-tie mass demotion buys cheaply, so the gate and the within-menu demotion cover disjoint wrongness, complementary rather than redundant.

Depth versus off-menu score margin. The stored analysis correlates error depth $G = s_{\text{lead}} - s_{\text{gold}}$ with $D = s_{\text{lead}} - \max_{j \notin \mathcal{O}} s_j$, on wrong items within each generator–suite case. Its implementation first forms $m = \max s_{\text{off}} - s_{\text{lead}}$ and then uses $-m = D$. Thus a positive correlation means deeper errors accompany a *larger* leader advantage over the best excluded candidate; it does not mean that candidate is pressing closer. The observed Spearman correlations are $+0.22$ to $+0.37$ on CSQA and $+0.20$ to $+0.39$ on RiddleSense, with positive intervals on $4/4$ and $4/4$ cases. Figure 6 shows all 21 cases. Point estimates are negative on all ARC and OpenBookQA cases; the largest point estimates are -0.11 and -0.03, respectively. Some OpenBookQA intervals cross zero, and neither suite has a positive interval. The five synthetic cases are mixed ($1/5$ interval-positive). These are descriptive, suite-dependent associations and do not derive the response to score demotion. Nor is the polarity a composition artefact: with legs scored by how much of their wrong mass sits beyond the pooled 1.67-nat depth median, the raw cross-leg rank correlation with the coupling is $+0.73$, but the suite-cluster bootstrap interval runs -0.06 to $+0.87$ and a threshold fit on four suites predicts the held-out suite's sign on only $10/21$ legs against the $3/21$ majority-label baseline — the coordinate is 89% between-suite, so suite identity, not depth composition, is doing the work (source `results/asm_geom_comp_law.json`; per-leg coupling, depth medians and wrong counts reproduce the geometry analysis). At the δ=3 endpoint the split hardens further: the rescue-surface difference reaches -87.0–-81.2 pp on CSQA — where the flag-free×non-rescued cell falls to zero items, so no odds ratio exists there (recorded as such, not smoothed) — and -52.1–-31.7 pp on riddle (-33.4–-14.9 pp on OpenBookQA), while ARC stays compatible with independence on both conditionals.

D.10 Limits of these diagnostic comparisons

The analyses above are measured on five suites of four-generator leg sets (21 legs); every wrongness claim holds for that audit corpus, and three scoping facts bound it. First, several cells are underpowered and recorded as such rather than smoothed: CSQA's flag-free wrong cell is four items (bracketed, -90.1–$+11.7$ pp), the deep-anchored suites' unheard cells are one and sixteen items (5.49/2.01 nats), and the synthetic concordance brackets one. Second, endpoints bite: at δ=3 CSQA's flag-free×non-rescued cell is structurally empty, so no odds ratio exists there, and ARC's unheard cell (42.1% of its wrong union) bounds what any flag-plus-demotion deployment can fix on that suite by construction. Third, the suite-level claims are five-cluster claims, not continuous laws: the depth-pairing polarity survives only as a suite property (cluster interval -0.06–$+0.87$, threshold transfer $10/21$), the union hearing is suite-dependent rather than flat, and the synthetic pools are directional evidence throughout (floors pool-substituted, geometry legs split $1/5$).

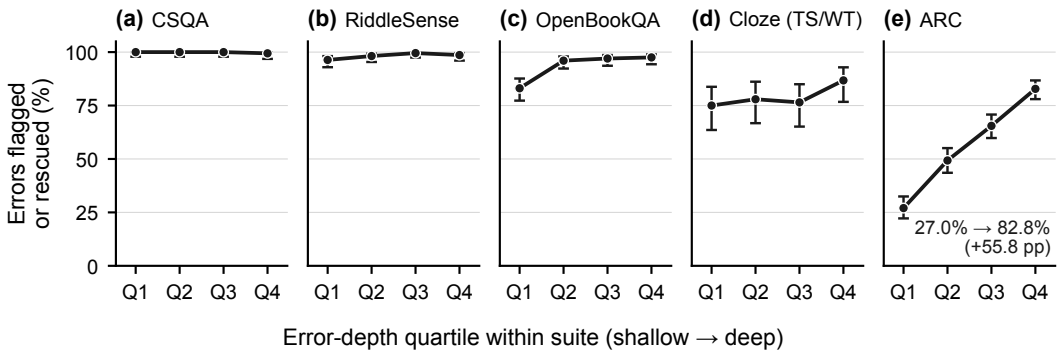

Figure 5: Fraction of erroneous items reached by the union of the off-menu flag and the $\delta = 1.5$ rescue diagnostic, grouped by within-suite error-depth quartile. Whiskers are recorded Wilson 95% intervals. Error depth and rescue are analysis quantities that use the gold answer; these curves do not describe a gold-free deployment policy. All five suites are retained, one panel per suite on a shared vertical axis (CSQA is CommonsenseQA); Cloze (TS/WT) pools the five controlled-cloze cases (TinyStories v3 for the four checkpoints, plus WikiText-103). The ARC gradient from the shallowest to the deepest quartile is labelled in panel (e).

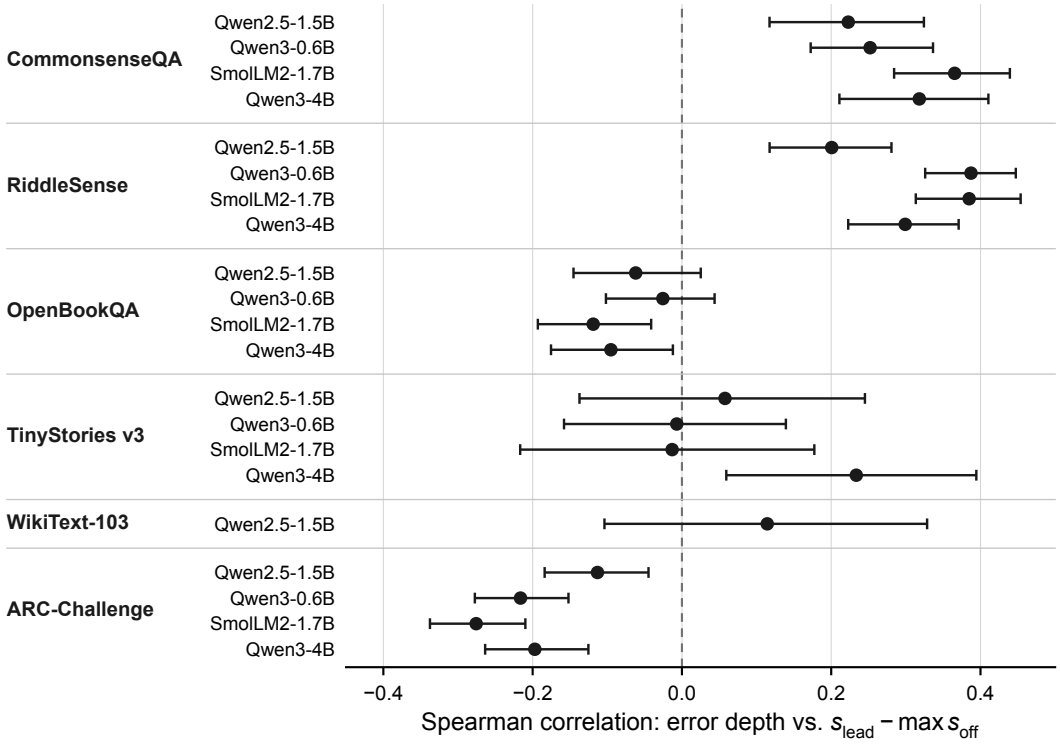

Figure 6: All 21 recorded score-geometry comparisons. Error depth is $s_{\text{lead}} - s_{\text{gold}}$ and the comparison margin is $s_{\text{lead}} - \max s_{\text{off}}$, with the sign shown in the axis label. Rows are grouped by suite; the five synthetic cloze cases are TinyStories v3 (four checkpoints) and WikiText-103 (host checkpoint only). Points are Spearman correlations among wrong items; whiskers are 2000-resample item bootstrap 95% intervals conditional on each checkpoint and suite. CommonsenseQA and RiddleSense intervals are positive; ARC and OpenBookQA point estimates are negative, with some OpenBookQA intervals crossing zero.

E SERVING POLICIES AND TRANSFER ACROSS GENERATORS

E.1 SERVING UNDER THE RE-MENU INTERVENTION

Running the full serving rule (MC-aggr, defined below) on the recorded draws gives the following results. At coverage 50%, served accuracy is $.999/.999$, served per-call reliability $.997/.999$, and the served set is mixed — flag rate $20.4/20.8\%$ among served (vs. global $30.7/37.1\%$). The audit layer's value is the menu-dependence meta-signal: among served items the flag predicts the pooled answer changing if w^* were served ($P(\text{flip} \mid \text{flag}) = .88/.88$ served vs. $P(\text{flip} \mid \text{clean}) = .03/.04$), so at coverage 50% roughly $21/21\%$ of served answers are menu-dependent, and $27/31\%$ at coverage 80% (ts v3 / wt v2; content-level flips on real re-menu draws). A downstream consumer thus gets, per served item: the calibrated probability (isotonic on the mass margin) and a menu-robustness flag — reliability-score and answer-stability are separately audited. *Confidence deflation* quantifies why the two signals must be separate: on flagged served items the within-menu per-call reliability overstates the covered-menu reliability by $\eta \approx .67$–$.78$ (real re-menu draws; ts v3 / wt v2, coverage 20–80), while clean served items deflate by at most $.11$; the net served-pool overstatement grows with coverage ($\approx .08/.16/.19$ at $20/50/80$, ts v3). The appropriate report for a flagged item carries the covered-menu reliability, not the within-menu number.

Combined with Section 2, one possible serving rule (MC-aggr) is **one forward; logit argmax answer; mass-margin serve; if $\tau \geq 1$, re-run with w^* served (on-demand candidate supplement) or report w^* as an alternate**; no repeated sampling, no gold at decision time.

Table 15: **Flag-augmented serving on real draws** (K=12; gold labels are used only in the fit and evaluation halves). Section 6 quotes the 50%-coverage risks of rows B and F. Flip is the rate at which the served pooled answer changes in content (compared by string) under the re-menu intervention. Risk is the realized gold-empirical error under the re-menu intervention, $1 - \rho^{\text{remenu}}$; the re-menu option at position 0 *is* the gold answer on both suites. The last column gives the coverage whose *fit*-half realized risk was at most 5%, with the evaluation-half realized risk in parentheses as the out-of-sample check ("none": no within-menu subset reached the target; fit floor $.153/.184$). Most fit-5% cells exceed 5% out of sample (evaluation $.064/.069$); the budget met out of sample is 2% on ts v3 (5.3% coverage, realized risk $.019$). † marks the one cell that over-selected on the fit half. For F1, risk ($-$) and the fit-budget coverage (n/a) are not reported. F^+ is a flag-conditional calibration reference: an isotonic fit of the re-menu gold frequency on the margin within each flag group (fit half). Deferral (F1) serves only the clean pool ($71/62\%$ actual coverage at an 80% target).

		At 50% coverage (served items)			Coverage at 5% fit budget
Suite	Rule	Flip	Risk $1 - \rho^{\text{remenu}}$	Acc	(evaluation risk)
ts v3	B within-menu	.197	.209	.997	none (fit floor .153)
	F deflated	.041	.104	.995	none
	F1 defer	.038	–	.992	n/a
	F^+ calibrated	.041	.109	.989	.247 (.064)
wt v2	B within-menu	.200	.227	1.000	none (fit floor .184)
	F deflated	.045	.114	.998	†.323 (.081)
	F1 defer	.038	–	1.000	n/a
	F^+ calibrated	.038	.108	.998	.191 (.069)

E.2 RE-ANSWERING, CONFIDENCE DEFLATION, AND ABSTENTION

What the flag should deliver. When the flag fires the downstream consumer has three candidate actions — serve w^* (the alternate), serve the original answer with deflated confidence, or abstain. The recorded evaluations compare these actions (`results/asm_flag_deliverable.json`): on the flagged/flipped served items w^* as the answer is (essentially) never correct ($P(\text{re-answer} \mid \text{flip}) = 0.00/.00/.00/.006$ on ts v3 / wt v2 / g2 / CSQA), i.e. replacing the answer with the displaced candidate — the menu-level competitor — loses on every configured suite; deferral (abstention to the clean pool) is correct on $.997/1.000/.993/.771$ respectively and is the flag's deliverable; the deflated rule's value is recalibrating the served confidence (η-gated $\hat{\rho}$), not answer correction. A

risk-aware consumer should therefore *quote* the deflated confidence but *decide* by abstention: defer flagged items, serve the clean pool, never serve the alternate.

E.3 Flags and the intervention on g2

The menu-flag layer also transfers: on TinyStories v3, g2's gold-free menu flag fires on 37.7% and realizes $P(\text{flip} \mid \tau \geq 1) = .913$ vs. .061 clean (content-level, real re-menu draws), $P(w^*$ becomes the pool answer $\mid \tau \geq 1) = .908$ vs. .045, precision .91, recall .92 (g1: .94/.10/.94/.81); the realized menu-exclusion drop is mean .277 (g1 .258), strictly monotone in $\log \tau$ quintiles (.004/.065/.231/.445/.638). All three layers — draw-free aggregation and calibration, the menu-exclusion intervention, and the menu flag — hold on two generators. *The serving layer transfers to the second generator as well.* (`results/asm_flagserve_e2e_g2.json`): the flag-deflated gate again halves served covered-risk at fixed coverage on g2's ts v3 — covered-risk @50 .272 → .160 (bootstrap $d = -.110$, 95% CI $[-.141, -.080]$, excludes 0), @20 .154 → .092 — and deferral cuts served content-level flips from .240 to .043 at @50 (CI $-.189$ $[-.229, -.147]$), with served pool accuracy $\geq .984$ and re-menu position-0 gold rate 1.000. The deflation factors reproduce ($\eta(\text{flag}{=}1) \approx .09$–.33 per margin bin vs. $\approx .78$–1.0 clean), so the deployment effect reported in Section 6 is not a 1.5B artifact either; the one difference is the size of the clean pool (g2 flag rate .377 vs. g1 .301, reflecting the 0.6B model's shallower preferences). The counterexample structure of Section C.1 replicates on g2 too: the strict content residual is 9.2% of flagged items (.908 take-over rate), and its plurality-stable items ($\approx .087$ of flagged) still lose .26 of per-call reliability ($40/49$ with drop > 0, exact sign-test $p = 5 \times 10^{-6}$), with the no-change rate falling from .42 in the threshold band $\tau \in [1, 1.4)$ (where w^* wins only half the time) to 0 for $\tau \geq 3$.

E.4 Cross-family validation (SmolLM2-1.7B-Instruct; non-Qwen)

To establish that the flag and intervention laws are not specific to a single model family, we run the identical one-forward extended-S masked pass on TinyStories v3 with SmolLM2-1.7B-Instruct (a 1.7B non-Qwen decoder; `results/raw/asm_offmenu_tsv3_smollm2.json`), and recompute the logit-space law in the paper's units (5-option-conditional gold reliability via ratio-invariant first-token restriction to the re-menu subset; `results/asm_offmenu_smollm2_consume.json`). The law reproduces at comparable scale: flag rate .284 (g1 .301), 5-option gold conditional .894 (g1 .879), predicted re-menu gold reliability .680, intervention drop mean $\Delta = .214$ ($P(\Delta{>}0) = .966$) — same sign and magnitude as g1's predicted .259 and real-draw .258 — with Δ *strictly monotone* in $\log \tau$ quintiles (.001/.019/.108/.318/.623; Spearman .90). The off-menu threat and its menu-exclusion law are therefore not a Qwen-specific artifact. The answer layer reproduces on real $K{=}12$ draws too: one-forward logit accuracy .9333 vs. the 12-draw poll .9293 ($\Delta = +0.4$pp, g1 $+0.7$pp on the same full-set basis) and vs. a single draw .885 ($+4.9$pp), with realized per-call gold reliability $\bar{\rho}^{\text{menu}} = .8935$ against the predicted 5-option conditional .8928; the "sampling most dangerous on hostage items" structure transfers (logit–poll disagreement .040 on flagged vs. .008 clean, flag rate .287). The real-draw flag and menu-exclusion layers complete on the re-menu half of the same job (n= 1480; `results/asm_g3_draws_consume.json`): realized re-menu gold mass 0.6816 vs. predicted 0.6804, and the realized drop $\bar{\Delta}^{\text{real}} = .2139$ matches the predicted 0.214 to 10^{-4} with per-item Spearman .81; the flag's displacement law reproduces ($P(\text{flip} \mid \tau \geq 1) = .779$ vs. .057 clean) and served flips are again one-sided ($P(\text{re-answer gold} \mid \text{flip}) = .023$ vs. .958 clean). Per-item realized drop sign is $K{=}12$-noise-limited ($P(\Delta^{\text{real}}{>}0) = .56$; the realized *mean* and ordering confirm the law). All three layers — answer, calibration, intervention+flag — hold on a non-Qwen generator with real draws; other architectures, scales, and $M \neq 5$ menus are not tested here.

E.5 Qwen3-4B on the controlled cloze menus

On the identical TinyStories v3 menus the larger 4B checkpoint (Qwen3-4B, from the same Qwen3 family as g2) receives the identical one-forward extended-S masked pass plus real $K{=}12$ masked draws on both the observed menus and a family-specific re-menu file in which the item's *own* strongest excluded candidate is supplemented (`scripts/asm_remenu_build_q4b.py`, deterministic; outputs `results/raw/asm_offmenu_tsv3_q4b.json`, `asm_qwen15_tsv3_q4b.json`,

	g1 Qwen2.5-1.5B-Ins	g2 Qwen3-0.6B-Base	q4b Qwen3-4B
answer acc, logit / count@12	.927 / .919	.893 / .886	0.904 / 0.902
serve acc @50, mass / count@12	.999 / .985	.991 / .980	0.997 / 0.965
serve deficit (mass−count@12) @50	.013	.011	0.032
quintile rho (asc. mass margin)	monotone	monotone	monotone
risk @50 / @90 (mass serve)	.001 / .039	.009 / .068	0.003 / 0.05
max cov @ risk $\leq 1\%$	65%	55%	62.2%

Table 16: Cross-generator reproduction of the draw-free structure on ts v3 (sources: `results/asm_g2_draw_calib.json`, `results/asm_q4b_draw_calib.json`).

`asm_qwen15_tsv3_remenu_q4b.json`, `asm_logit_tsv3_q4b.json`). *Flag and intervention.* The unit flag fires on 31.1% and realizes $P(\text{flip} \mid \tau \geq 1) = 0.936$ vs. 0.04 clean, with $P(w^*$ becomes the pool answer $\mid \tau \geq 1) = 0.929$ vs. 0.034 — precision 0.929, recall 0.925 (g1 .94/.81, g2 .91/.92: at the top of the three Qwen legs' recall range). The realized re-menu drop has mean 0.254 (g1 .258), monotone in $\log \tau$ quintiles (0.000/0.005/0.083/0.435/0.746), and prediction matches realization tightly: predicted drop 0.2532 vs. realized 0.2525 (mean abs. err. 0.049, per-item Spearman 0.83; predicted re-menu gold mass 0.637 vs. realized 0.636), with flips one-sided as before (`results/asm_q4b_flag_validate.json`, `results/asm_q4b_draws_consume.json`). *Answer and serve dominance.* On real draws the one-forward logit argmax wins (0.904 vs. count@12 0.902 and single-draw 0.885; g1 .927/.919/.872), count–logit disagreement 0.025; the zero-draw mass-margin serve dominates the count serve at cov50 (0.997 vs. 0.965; deficit 0.032), per-call calibration is quintile-monotone (0 inversions), and the risk–coverage structure holds (risk@50 0.003; max coverage at risk $\leq 1\%$ 62.2%, g1 65%/g2 55%) (`results/asm_q4b_draw_calib.json`). *Deflated deployment gate.* The end-to-end deployment cell replicates (`results/asm_flagserve_e2e_q4b.json`): the flag-deflated gate cuts served covered risk at fixed coverage (@50 0.18 $\rightarrow$ 0.066, bootstrap $d = -0.099$, 95% CI $[-0.134, -0.067]$ excludes 0; @20 0.154 $\rightarrow$ 0.057), deferral cuts served content-level flips from 0.179 to 0.029 at @50 (CI -0.147 $[-0.184, -0.115]$), the deflation factors reproduce ($\eta(\text{flag}) \leq 0.17$ per margin bin vs. $\eta(\text{clean}) \geq 0.92$), the re-menu position-0 gold rate is 1.000, and served flips stay one-sided ($P(\text{re-answer gold} \mid \text{flip}) = 0.006$ vs. 0.939 clean). The flag results are therefore not specific to the 1.5B host: they hold from 0.6B to 4B, across two Qwen generations and one non-Qwen family.

E.6 Cross-generator numerical summary

Table 16 summarizes the draw-free structure on ts v3 for g1, g2, and Qwen3-4B.

E.7 Serving under deep under-inclusion (CSQA cross-benchmark)

The four serving rules of Section 6 run on the single-token CSQA subset (real $K=12$ draws, real re-menu draws, same OOS protocol). *Identity join.* Joining the single-token subset to the full-validation run by *numeric row index* would match 0 benchmark IDs: of the 941 numeric overlaps only 103 items share a real CSQA id, 99 of them with re-ordered option lists (aligned by answer string by the identity join `scripts/asm_csqa_identity_join.py`). The matched 103-item subset replicates the mechanism (`results/asm_flagserve_csqa_{tau,treal}_identity_v2.json`) but is underpowered for serving deltas; we therefore also report a full-941 *identity-safe* serving estimate that reads *only* the single-token artifacts (`results/asm_flagserve_csqa_941_v3.json`): because every array is keyed by the same single-token construction index, no cross-run join exists at all. *Margin definition.* The single-token candidate masses store the 21 extended-S entries (5 menu options plus 16 off candidates); the top-2 gap over all 21 is not a within-menu margin, because on 678/941 items the over-the-set leader lies *outside* the menu. The serving policy (`scripts/asm_serving_policy_v3.py`) therefore uses the in-menu conditional margin over the first-five renormalized masses, score-descending with a fixed seeded tie-key, with the same selection in Table 3, the fitted frontier, and the resample bootstrap (duplicates retained). Replaying the variant combinations shows that the covered-risk difference between the over-the-set and in-menu margins comes from the margin input definition, not the tie/isotonic policy. The single-token 941 observed and re-menued draws are unaffected by either issue; the two flag bases agree on the matched subset

(both give flag rate .757; mechanism below). The deflation mechanism is visible ($\eta \approx 0$–$.25$ on flagged vs. .33–1.0 clean), and the re-menu option at position 0 equals the gold string on every item (100% string check in both estimates). The flag-aware laws transfer in form, with limitations:

The within-menu gate is at its worst here. On the full-941 estimate (fit 470 / eval 471), serving the top half by the in-menu conditional margin (baseline B, cov 50%, 235 served): served pool accuracy is .83, the served re-menu pool answer is gold on .33, 60% of served answers change content under the re-menu (all correct→wrong, P(re-answer gold | flip) = 0.0 vs. .84 clean), and the served covered-gold-error is .65 — deep under-inclusion persists and the within-menu serve still delivers mostly hostage answers. *The flag reduces — but does not halve — covered risk at full scale.* With the threat-real flag (the masked-mass τ of the single-token forward pass) in the gate, the deflated rule F at cov 50% lowers served covered-gold-error to .61 (.610 vs. .652), cuts served flips (.481 vs. .604) and raises the served re-menu pool gold rate (.33 → .41), at a served-accuracy cost (.830 → .783; 95% CI of the accuracy difference $[.00, +.10]$); the fixed-top-k eval-item bootstrap (2000 resamples, frozen-fit calibration: the intervals are eval-item resampling only, fitting uncertainty is not included) gives an $F-B$ covered-risk difference of -0.042 with 95% interval $[-.067, -.018]$, excluding 0 but about half of the -0.089 obtained with the over-the-set margin. Deferral ($F1$) cuts served flips on its reachable cell (-0.41 $[-.50, -.32]$ at cov 20%; at cov 50% it serves the clean pool at .78/.69 acc/re-acc with 12% flips), and the clean pool caps its coverage at $\approx$ 28% (versus 71/62% on the synthetic suites), so deferral withholds roughly 70% of items. *Consequences.* (i) The served risk reduction at 50% on the identity-safe full-941 basis is significant but modest ($F-B = -0.042$, 95% CI $[-.067, -.018]$); the *tight-budget* frontier stays degenerate regardless of gate ($\leq$ 1–2% budgets unreachable; the $\leq 5\%$ budget only at $\lesssim 1\%$ coverage). (ii) The reduction is margin-input dependent and sample limited: with the multi-token length-normalized margin the matched 103-item subset cannot resolve it ($F-B = -.07$, CI $[-.16, +.02]$), while the in-menu margin at n=941 excludes zero; both use the same flag and draws, so the gap is margin basis and sample size, not the flag. (iii) *The full-941 panel is seed-stable:* re-running the serve fit under four further fit/eval split seeds leaves every bootstrap CI for the $F-B$ covered-risk reduction excluding 0 (`results/asm_flagserve_csqa_941_seed{101,2025,7777,424242,31337}.json`), so the modest-significant result is not an artifact of one split or one η fit. The spread law is quantified cell-wise in `results/asm_multiseed_csqa941_parity_spreadlaw.json`: against the binomial split-sample reference (a half-split of 471 out of 941 items off a fixed pool, finite-population corrected; expected five-seed max–min = $2.326\,\sigma$), the within-menu rule is exactly at scale (sd ratios 0.72–0.90; spread ratios 0.83–1.05), while the two flag-side quality rates carry a mild extra slack of the per-seed η/gate refits (sd ratio 1.27; spread ratios 1.52/1.59) — coverage- and risk-level cells stay at scale (ratios 0.70–1.00/0.87–1.21), and the flag rate itself is split-invariant by construction (draw-level). *Consumer utility of the flag.* Because the alternate action is ruled out in the measured suites (serving w^* is never correct on flagged flips), we quantify the utility of abstention. On the identity-safe full-941 CSQA panel, across the three coverage budgets $\{20, 50, 80\}$% the flag-deflated gate averts a mean $\approx$ .077 covered-gold-error (B.641→ F.564), and at cov 50% the 235 served items lose 153 answers under B versus 143 under F (10 wrong answers averted per 235 served); the deferral action serves the $\approx$ 28.5% clean pool at .78/.69 acc/re-acc with served flips 12% (a $\approx 5\times$ reduction from the baseline's 60%). So the flag's value is *abstention plus recalibrated confidence*, not replacement: a coverage-limited consumer who can absorb a 35–67% clean fraction (CSQA vs. ARC) buys a $5\times$ flip cut and a modest but significant covered-risk reduction, the modest payoff the measurements support. *Mechanism.* On the identity-matched 103 CSQA items the likelihood-τ and threat-real bases *coincide* — $\mathrm{corr}(\log \tau_{\mathrm{LL}}, \log \tau_{\mathrm{real}}) = 1.0$, full agreement (no only-LL / only-real groups), and displacement AUC .978 for both (`results/asm_csqa_llflag_mechanism_identity_v2.json`). A numeric-index join would instead suggest orthogonal bases (AUC .49); that reading is an artifact of the join, not a property of the generators: a length-normalized likelihood ratio and a masked-mass ratio are the same ranking on single-token items. *Full-scale same-identity match.* Lifting the match off the 103-item subset, a fresh likelihood read over all 941 identity-safe single-token CSQA items (same 16 off candidates as the masked-mass excluded-candidate scores; `results/asm_csqa941_ll_consume.json`) gives $\mathrm{corr}(\log \tau_{\mathrm{LL}}, \log \tau_{\mathrm{real}}) = 1.0$ and $P(\mathrm{flag}_{\mathrm{LL}}$ agree $\mathrm{flag}_{\mathrm{real}}) = 1.0$ (both $P(\tau \geq 1) = .7205$), identical LL- and masked-argmax accuracy (.660/.660), and real-draw serving deltas of the same sign under both bases (F lowers served covered-gold-error at cov 50% on each; deferral serves the identical clean pool at 28.5% coverage). On this single-token basis the two readers are the same instrument at full scale — which is what

the 103-item correlation predicted and is the scope the multi-token serving cell (above) transfers by deferral. The flag's displacement precision replicates on the full set ($P(\text{flip} \mid \text{flag}) = .94$ vs. .08 clean; realized $\rho_w - \rho_c = .50$), consistent with Section 4. *Contamination handling.* The full-941 estimate pools the 838 CSQA training rows with the 103 validation rows. On the validation rows alone the flag and intervention panel reproduces with no degradation — flag rate .757, $P(w^* \mid \text{flag}) = .949$ vs. .080 clean, realized drop 0.4183 vs. predicted 0.4193 (Spearman .95), pool-answer collapse under re-menu -48.5pp (the train rows: $.716/.958/.385/-43.9$; `results/asm_csqa_validation_split.json`) — so neither the under-inclusion nor the drop law is a training-row artifact. (The split bounds the benchmark's own train/validation protocol; CSQA predates the generator, so the split alone cannot rule out that validation rows appear in the generator's pretraining corpus — the claim is "not a benchmark-training artifact", not "not in any training material".) Serving deltas remain the full-941 estimate only because n=103 cannot resolve a serving CI. Serving gains on CSQA therefore equal whatever the single-token flag carries; whether a cheap basis separates from the masked-mass threat on *multi-token* answers remains open (Appendix I), answered structurally by the negative span-level result and pragmatically by the likelihood-basis deferral transfer. (Cost note: the twelve cached single-token draws reuse one forward distribution; full-menu likelihood candidate scoring is not decomposed into forward counts by the stored arrays, so no "12× forward savings" claim is made on this cell.) The synthetic-suite serving laws are therefore not an artifact of easy menus, but neither do they transfer for free: flag informativeness and achievable coverage are the transfer currencies. *Serve-layer laws at the* 40–65% *base-accuracy regime.* The synthetic serving cells sit at 93–95% base accuracy; coverage-limited serving operates where reliability is lower, and we ask directly how the serve-dominance gap and the quintile calibration behave there. On the same identity-safe full-941 basis (base pooled accuracy .649; `results/asm_csqa_serve_transfer_q4.json`), the *one-forward mass-margin vs. K=12 count-margin* dominance transfers at the tightest coverage: at 20% the mass gate serves .979/.915 pooled-accuracy and .500/.426 re-menu gold-answer rate (bootstrap difference in served accuracy +.063, 95% CI [0.00, .128]), narrowing to indistinguishable at 50/80% coverage (where both menus' deep under-inclusion saturates the served risk) — the zero-draw advantage over the poll persists where a server is most selective, and vanishes where selectivity is diluted. The *margin-quintile calibration* is inversion-free at this regime too: pooled observed accuracy rises .329/.462/.648/.756/.979 across the five margin quintiles and re-menu gold reliability .051/.066/.181/.233/.505 (0 inversions on both, Spearman 1.0) — the monotone law survives at low base accuracy, with the caveat that the *top* quintile's re-gold reliability is only $\approx .5$: the calibration curve is right, but deep menu under-inclusion caps the ceiling that calibration can certify.

Span-level masked reads do not repair the multi-token flag (negative result). We attempted to generalize the extended-S masked-mass threat to multi-token answers via span-level masked reads, and implemented it on ARC (M=4 multi-token) in two variants. (i) *Fixed-width union mask* (`results/asm_arc_span_consume.json`): per span position, mask over the distinct option/off tokens present there, multiply the conditional masses along each option's span, and compare the off-candidate span scores to the leader option span score. The result is a *systematic over-flag*, not a repaired basis: $P(\tau_{\text{span}} \geq 1) = .885$ vs. the LL-ratio's .333, answer accuracy .456 (no better than the LL reader's .469); the mechanism is structural, not statistical — the per-position mask treats the 16 single-token off candidates and the M option tokens as equal alternatives, so an off token beats the leader's span score on $10.4/16$ candidates per item and the off side collects $\approx 16\times$ the mass by construction. (ii) *w^*-own-span single comparator* (the displaced candidate's own 2-token continuation masked at each step, τ = single strongest off span vs. leader option span; `results/asm_arc_span2_consume.json`): this is the repair that the diagnosis of variant (i) calls for, and it removes most of the $16\times$ inflation — yet the flag rate stays far above the LL basis (.808; validation .858 / test .796 vs. LL .393/.318), the masked-read answer is *less accurate* than the LL answer (.416 vs. .469; disagreement on 25% of items), and τ_{span2} does not rank the LL answer-correct weak label (corr=−.03 vs. LL's +.02). The residual over-flag is structural and uniform across answer length (span2 rate .65–.93 at gold lengths 1–11, where the LL ratio *declines* with length, .44 → .14): a greedily selected continuation out-masses a fixed answer string in per-step masked reads at *every* length, because the per-step masks are not length-commensurable and the off span is generator-optimal by construction. The multi-token flag basis on ARC is therefore the length-normalized likelihood ratio itself ($P(\tau \geq 1) = .333$, established), not a masked-read repair; on the matched CSQA items the LL-ratio and masked bases coincide, so no separate repair is motivated there. The

masked-read path is thus a second, independent negative result; a same-identity full-CSQA matched likelihood serving run and a real-draw span validation remain open (Appendix I).

Plausibility audit of the excluded candidate (arbitrating the "defective menus" vs. "weak model preference" reading). We ask whether the 65–76% under-inclusion rate means the human menus are defective or the model's single-token QA preference is weak; a data-driven audit (`results/asm_csqa_plausible_disp_v2.json`) calls the re-menu candidate w^* (the option the re-menu added) plausible iff its string is itself the gold answer of some CSQA question in the local item files. On the single-token subset only $P(w^* \text{ plausible}) = .372$: most off-menu candidates are *not* plainly reasonable answers by the benchmark's own label standard (top offenders include shallow tokens such as *kitchen*, *home*, *library*); $P(\text{plausible} \mid \text{flag}) = .38$ (flag = the single-token set's own threat-real flag, read on the single-token items themselves rather than through a numeric row index into the full-validation multi-token run; the plausibility rate does not depend on this choice). Under-inclusion is insensitive to the distinction ($P(\tau \geq 1) = .721/.731/.714$ all / plausible-w^* / not-plausible-w^*), and the displacement law is unchanged on plausible candidates ($P(\text{flip} \mid \tau \geq 1,\ w^* \text{ plausible}) = .934$ vs. .957 not-plausible). So the mechanism — the presented menu under-covers the generator's own continued-preference – is robust, but the *"accuracy inflation"* reading deserves a qualification: the displaced-to answer is frequently a label-off-plausible continuation of a 1.5B model, so the serving stakes of under-inclusion are a "model would report a worse-than-leader preference" effect, with the inflation amount dampened by the low plausibility rate. *Operationalization sensitivity.* The "no smaller on plausible (w^*)" reading is robust to how the label is drawn (`results/asm_csqa_plausible_sens.json`): using the union gold vocabulary (37.2% of w^* pass), the single-token-only golds (32.8%), or only gold strings appearing ≥ 2 times locally (26.7%; a 323-word vocabulary), the pooled-answer displacement on the plausible-w^* subgroup is $-46/-47/-45$pp — in every case *at least* the overall -44pp — so the conclusion does not ride on the specific label rule. *Joint flag∩plausible cell.* The sharpest single "inflation" quantity is the measured-accuracy displacement on the items that are *both* flagged ($\tau \geq 1$) and plausibility-audited: this joint set is 27.2% of the suite (n=256) and its pooled-answer accuracy moves .641 → .035 (-60.5pp) when the single excluded plausible candidate is supplemented — on almost every jointly-flagged-plausible item the answer is set by menu completeness, which is the defensible inflation number a consumer can attribute to the menu. *Item-specific validity and error-repair significance.* The key remaining objection is whether the excluded candidate is ever a *valid answer for the particular item*, which would make the displacement a genuine menu-omission error instead of a model-preference artifact. The strongest audit available without external annotation (`results/asm_wstar_gold_validity.json`; same identity-safe per-item artifacts, seeded tie-break) compares w^* against the *same item's* gold answer: $P(w^* \text{ is the item's gold}) = .000$ on both ts v3 and CSQA, both unconditionally and conditional on the flag ($\tau \geq 1$), and the gold answer is present in every presented menu (0 omissions) — the excluded candidate is *never* the omitted correct answer, on synthetic or human menus. Consistently, the supplement does not *repair* wrong pooled answers: among flagged items whose observed pooled answer is wrong, the re-menu answer returns to gold on $2/65$ (ts v3) and $1/265$ (CSQA). This is the separation the design demands: the flag marks menu-dependence of the generator's preference (an answer-stability audit — abstain/deflate are its actions), not that the benchmark omitted a legitimate competing answer (which would license serving the alternate). The "accuracy inflation" quantity is a menu-sensitivity measurement, not a claim that the gold answer was missing. The audit therefore keeps the mechanism claim and softens the inflation one on the CSQA item family.

Plausibility-restricted serving. We ask whether the serving protocol still pays once w^* is restricted to plausible candidates. A numeric-index join of the same single-token items to the full-validation run yields gate numbers (.703 → .627, d=−.080) computed on misidentified items, so we do not use them; the identity-matched serving sample (n=103) is too small to split by plausibility. What survives identity-safely is the *measured-accuracy* half (`results/asm_csqa_plausible_disp_v2.json`): pooling over the re-menu on plausible-w^* items (n=350) displaces measured accuracy *at least* as much as on the full set — observed .651 → .194 re-menued (-46pp), vs. -44pp overall and -43pp not-plausible. So the audit's displacement reading stands (the measured-accuracy shift is not carried by implausible shallow continuations), while no serving claim is made within the subgroup (it would require a same-identity full-941 matched likelihood run).

E.8 CSQA INTERVENTION AND SERVING ACROSS FOUR CHECKPOINTS

The full g1 analysis chain of Appendix E.7 — family off-menu scan (own top-4096 next-token candidates, good-filtered), deterministic re-menu build (gold, top-3 in-menu leaders, w^*), real $K{=}12$ draws on both menus, logit-condition prediction, and the identity-safe serving analysis — is rerun unchanged for the three non-g1 families (fp32, seed 20260903; `results/asm_csqa941_remenu_consume_4fam.json`, `results/asm_flagserve_csqa941_v3_*.json`). Table 17 quotes the flag and intervention layer; Table 3 in Section 6 quotes the gate. (i) *The flag is frequent on this suite and saturates for g2 and* SMOLLM2: $P(\tau{\geq}1)$ is 0.721 (g1) vs. 0.773 (Qwen3-4B), 0.947 (g2), 0.946 (SMOLLM2) — while $P(w^* \mid \tau{\geq}1)$ is .94–.99 and recall .97–.99 on all four, i.e. the deep families' strongest off-menu content outscores the human menu leader on almost every item and almost always wins the re-menu pool. (ii) *The intervention law holds per item on every family*: realized and predicted re-menu gold drops agree at the mean level within 10^{-4} (g1), 0.003 (g2/SMOLLM2) and 0.012 (Qwen3-4B) with Spearman .929–.955, and measured pooled accuracy shifts by $-43.36/-47.61/-41.76/-48.54$pp (g1/Qwen3-4B/g2/SMOLLM2). (iii) *The deflated gate's measured utility is set by flag saturation*: at g1's 0.721 flag rate the gate cuts served covered risk $0.652 \to 0.61$ at cov50 ($-0.042[-0.067, -0.018]$), while at the g2/SMOLLM2 rates ($0.947/0.946$) the deflated side retains only a 0.055-cap clean pool and the comparison is a wash ($+0.011[-0.002, +0.025]$, $+0.011[-0.004, +0.024]$) — the gate has nothing left to serve when nearly every item flags (Figure 4). The mechanism is scarcity, not a fading deflation carrier: the re-menu reliability of flagged items is zero at the ρ_c/ρ_w median on every family's fit half, so the deflation factor applied to flagged eval items is strong on all four (.021/.003/.000/.008), and the same carrier holds on both re-randomized splits ($12/12$ legs at a zero median; effective flagged-eval deflation $\leq .021$ everywhere; `results/asm_gate_eta_splits.json`) — yet at the saturated flag rates the cov50 served pool is $\sim 90\%$ flagged whichever ordering is used (only a 0.055-cap clean top-up exists), leaving no exploitable differential; at g1's lower flag rate the same carrier routes service through a $1{-}0.721$ clean pool and the cut appears. Measured pool composition says the same in counts: the cov50 served pool's flagged share falls $.575 \to .430$ at g1 and $.604 \to .523$ at Qwen3-4B, but only $.898 \to .889/.898 \to .893$ at the saturated pair. Qwen3-4B sits between: $-0.016[-0.037, +0.003]$ ($0.68 \to 0.66$, cap 0.238). Split robustness: re-randomizing the fit/eval split, tie key and bootstrap seed (two alternate offsets, same draws) leaves every family in its measured class — g1 stays a cut with all three CIs strictly negative ($-.042/-.038/-.036$), Qwen3-4B's cut stays partial with all CIs touching zero ($-.016/-.021/-.018$), and the saturated pair's signed difference stays within $.001/.019$ with split-dependent CI sign (.011/.016/.001/.011/.019/.015) — the small difference there is a statement about magnitude, not about any single fit (alternate splits: `results/asm_flagserve_csqa_941_v3_splt{1,2}.json` for g1 and `results/asm_flagserve_csqa941_v3_{g2,smol,q4b}_splt{1,2}.json` for the three other checkpoints). Draw-seed robustness extends the statement from three fits to three independent draw sets: re-drawing both menus at two further seeds (20260904/20260905; same re-menus, runner, $K{=}12$ and protocol) reproduces the class structure not only on the base split ($-.042/-.059/-.047/-.016/-.028/-.018/.011/.011/.012/.011/.014/.011$; g1 3/3, saturated pair $4/4$ in the reference band) but across the full $3{\times}3$ draw-seed-by-split grid (nine cells per family; coupled partitions, so a column compares draws on an identical split): g1 is a strictly negative cut on every cell ($9/9$), Qwen3-4B is negative on every cell ($9/9$; seed-20260904 reaching a strictly negative CI at split 0), and the measured class holds on $35/36$ cells — the single deviation is one SMOLLM2 cell at seed 20260905, split 1: $-.0004$ with CI $[-.0154, .0142]$, recorded as a class deviation; the estimated effect remains small (matrix: `results/asm_gate_seed_split_matrix.json`; readout: `scripts/asm_gate_matrix_readout.py`). The cross-draw dispersion of the cell centers tells the same story in one number per split (`results/asm_gate_crossdraw_spreadlaw.json`): against the pooled within-cell bootstrap halfwidth (noise expectation $0.864{\times}$ mean halfwidth for three independent draws of an invariant effect), g1 stays at 0.29–0.87 and Qwen3-4B at 0.03–0.64 on every split, while the saturated pair sits at noise scale except the exact split where its deviation lives — split 1 carries 1.20 (g2, class intact) and 1.35 (the documented SMOLLM2 deviation cell itself): the single stated class deviation is precisely the single spread outlier of this analysis. The deviation cell's collapse decomposes exactly (`results/asm_smol_ds05_split1_decomp.json`): at the coupled partition — margins, flags and tie keys draw-invariant by construction — the point move $0.0187 \to -0.0004$ (collapse 0.0191) splits additively into a 0.0159 served-set churn term (15 of 205 F-side and 11 of 205 B-side served items swapped under the redrawn menus) against a 0.0033 per-item remenu-mass drift, while the within-

band comparator draw moves just 0.0061 (0.0069 churn partly offset by a -0.0008 mass drift; 2 of 205 F-side swaps): the class deviation is a served-set composition effect at the within-cell noise scale, not a systematic effect of the menu redraw. Magnitude, unlike class, differs on the cut families: their $F-B$ magnitudes exceed the reference three-split band on the new seeds *in the favorable direction*, and the readout records the band criterion as failing there (`results/asm_gate_drawseed_readout.json`; analysis script: `scripts/asm_gate_drawseed_consume.sh`; per-checkpoint results: `results/asm_flagserve_csqa_941_v3_ds{04,05}.json` for g1, `asm_flagserve_csqa941_v3_{g2,smol,q4b}_ds{04,05}.json` for the rest). Scope: SMOLLM2's build covers $820/941$ items (121 dropped by the single-token round-trip rule, the same rule that dropped none for the other families): $85/121$ fail on a top-3 distractor token and $36/121$ on the gold token (zero on w^*), and the dropped items are threat-profile-neutral against the kept pool — flagged-regime membership .950 vs. .946, median threat 28.3 vs. 40.1 — so the gate-relevant regime is preserved by the restriction (composition: `results/asm_csqa941_smol_dropped_profile.json`, `asm_csqa941_smol_dropped_decompose.json`). Qwen3-4B's deepest tail saturates the 6-decimal storage floor ($18/941$ items at zero stored in-menu mass); a scan at 10^{-10} precision (same runner, seed and top-k) recovers 17 of these and confirms one item (index 767) whose entire in-menu mass is zero already at fp32, so the within-menu-margin gate analysis excludes that single item symmetrically from both gate arms ($n = 940$; `results/asm_flagserve_csqa941_v3_q4b.json` records the exclusion; every other family's artifact is exact at the quoted digits).

Table 17: **Flag and intervention statistics for the four checkpoints on single-token CommonsenseQA** (941 items, 820 for SmolLM2-1.7B as described above; five-option menus; all items). The flag rate is $P(\tau \geq 1)$. Precision is $P(w^* \text{ wins} \mid \text{flag})$, the rate at which w^* becomes the pooled 12-draw answer on the altered menu among flagged items, and recall is $P(\text{flag} \mid w^* \text{ wins})$. Δ_{real} and Δ_{pred} are the mean realized and logit-predicted drops in gold mass after the intervention, and ρ_s is their per-item Spearman correlation. Acc. change is the measured change in pooled 12-draw accuracy after the item's own w^* replaces a distractor (negative values are decreases). Source: `results/asm_csqa941_remenu_consume_4fam.json`.

Checkpoint	Flag rate	Precision	Recall	Δ_{real}	Δ_{pred}	ρ_s	Acc. change (pp)
Qwen2.5-1.5B (host)	.721	.944	.967	.388	.389	.937	-43.36
Qwen3-4B	.773	.988	.977	.456	.452	.976	-47.61
Qwen3-0.6B	.947	.983	.991	.351	.346	.929	-41.76
SmolLM2-1.7B	.946	.988	.992	.404	.404	.933	-48.54

The four-checkpoint gate comparison at 50% coverage (within-menu baseline B, flag-deflated gate F, paired bootstrap $F-B$, and clean-pool cap) is reported in Table 3 of Section 6. It uses the identity-safe serving analysis, seeded fit/evaluation halves, and 2000 bootstrap resamples conditioned on the fixed fit half, with each checkpoint's own artifacts and SMOLLM2 on its 820-item build (sources: `results/asm_flagserve_csqa_941_v3.json` for g1 and `asm_flagserve_csqa941_v3_{q4b,g2,smol}.json`).

Full coverage curves. Figure 4 plots all recorded coverage points. At 50% coverage, the reconstructed curves lower risk for the host and Qwen3-4B, while Qwen3-0.6B and SmolLM2 show small increases. Their empirical curves are not identical. The shaded ranges span three fit/evaluation partitions, including their tie-key changes, and are not confidence intervals. The per-cell bootstrap intervals above describe their recorded tie-rule realization; they need not equal differences of separately reconstructed curve values to the last rounded digit. The checkpoint-specific coverage limits and negative differences remain part of the calibration result.

F ESTABLISHED BENCHMARKS: EFFECT SIZES AND FAILURE BOUNDARIES

F.1 FULL-BENCHMARK GENERALIZATION (MULTI-TOKEN ANSWERS)

In the multi-token comparison, we score every option by its *length-normalized sequence log-likelihood* (one prefix forward plus a batched forward over partial-option sequences), select the argmax answer, and construct off-menu candidates as above; ratios are normalization-invariant, so

$\tau = \exp(\max_{\text{off}} \ell - \max_{\text{opt}} \ell)$ extends the same-basis ratio to unequal-length options. On the full CommonsenseQA validation set (N=1197, human-written distractors, multi-token answers): MC-likelihood accuracy is 0.610, and under-inclusion $P(\tau \geq 1) = 0.649$ — the generator's single most-preferred content next-word outscores the best presented answer option's normalized likelihood on 65% of items (single-token-subset reference $0.72/0.77$: the one-forward reads of the gold-free τ and of its gold-referenced gain variant on the 941; `results/raw/asm_offgen_csqa.json`). Accuracy rises with answer length ($0.54/0.65/0.74/0.92$ for gold token lengths $1/2/3/4$), i.e. the model struggles most on the shortest answers — the regime the single-token real-draw anchor covers. The mechanism—human-distractor menus under-cover the generator's preferred answer—is not confined to single-token answers.

F.2 Second established suite: ARC-Challenge

The same multi-token likelihood protocol runs on AI2 ARC-Challenge (Clark et al., 2018) (M=4, human-written science-choice distractors, validation+test N=1460), using the generator-native off-candidate construction (`results/asm_arc_consume.json`). MC-likelihood accuracy is 0.469 (validation 0.485 / test 0.465), with a mild length trend ($0.41/0.44/0.46$ for gold answer lengths $1/2/3$ tokens); the under-inclusion rate is $P(\tau \geq 1) = 0.333$ overall (validation 0.393 / test 0.318) — the generator's own off-menu content continuation beats the presented leader on a third of ARC items. The suite calibrates the magnitude claim: ARC's under-inclusion ($\approx .33$) sits *with* the synthetic-cloze family (.30–.37) and far *below* CSQA's (.65) and OpenBookQA's ($\approx .76$; Appendix F.3) — deep under-inclusion is not universal: it is deep on CSQA and OpenBookQA, shallow on ARC. The mechanism (the leader is not the generator's preferred continuation, measured in the generator's own likelihoods on M=4 menus) transfers to the second and third established suites.

F.3 Third established suite: OpenBookQA

OpenBookQA (Mihaylov et al., 2018) (M=4, human-written science-commonsense answer choices, validation + test N=1000, both held-out, obtained from the Hugging Face hub) runs the same multi-token likelihood protocol with generator-native off candidates (`data/asm_forced_obqa_all.jsonl`, `results/asm_obqa_stakes_v2.json`): MC-likelihood accuracy is 0.424 (validation 0.390 / test 0.458), and the under-inclusion rate is $P(\tau \geq 1) = 0.755$ (validation 0.744 / test 0.766) — deep, comparable to CSQA (.65) and far above ARC (.333), so deep under-inclusion is *not* a CSQA quirk of human-written distractors. The three-family stakes on the identical 1000 menus (Appendix F.5): own shifts $32.2/15.5/16.3$pp (g1/SmolLM2/g2, bootstrap CIs $\pm \leq 3$pp) against an 8.7pp largest pairwise gap, all three exceeding the gap (joint CIs exclude 0); the family ranking reproduces CSQA and ARC (g1 > SmolLM2 > g2) with 0 item-level pair-relation reversals, and the collapse identity holds with 0 violations (twelve family×suite cells in total). Serving-layer and count-poll anchors on OpenBookQA are not measured (scope: the stakes / under-inclusion cells only).

F.4 Fourth established suite: RiddleSense

RiddleSense (Lin et al., 2021) (M=5, human-written riddle options, labeled dev partition N=1021; the official test keys are not publicly released) is a deliberate stressor for the displacement law: 86% of the golds and 82% of the options are single tokens, so the generator's single-token off-menu pool competes *directly* with the answer set (`data/asm_forced_riddle_all.jsonl`, `results/asm_riddle_stakes_v2.json`). The under-inclusion rate is $P(\tau \geq 1) = 0.82$ — the deepest of the four established suites — and the three-family stakes on the identical 1021 menus reproduce the ranking (g1 > SmolLM2 > g2) with 0 item-level pair-relation reversals, and under-inclusion is deep ($P(w^* \mid \text{re-menu answer}) = 0.8198/0.5563/0.7169$, g1/SmolLM2/g2): own shifts are $36.2/18.5/19.8$pp (bootstrap CIs $\pm \leq 3.1$pp) against the 20.8pp largest pairwise gap — the largest family gap on any established suite. The shift dominates the gap only for the deepest-under-included, most able family: $+15.4$pp (joint 95% CI $[12.4, 18.7]$, P=1.0), while the two weaker families land *on* the gap ($-1.0/-2.3$pp, joint CIs $[-5.7, 3.7]/[-6.3, 1.8]$, P=.348/.124) — unlike ARC their shifts are not significantly below it: deep under-inclusion plus a near-chance weaker family (acc $0.2752/0.331$) yields an accuracy-weighted shift that sits within error of a large family gap. The closed-form boundary $P(w^* \mid \text{re-menu answer}) > \text{gap/acc}$ still selects the crossing

family on all 12 cells (`results/asm_stakes_shiftlaw_pooled.json`); the hostage fractions (shift/acc) are 0.75 (g1), 0.559 (SMOLLM2) and 0.719 (g2) — deep displacement despite the moderate absolute shifts, because the weaker riddle families are near chance — and the item-level collapse identity holds with 0 violations.

F.5 MENU-INDUCED ACCURACY SHIFTS VERSUS MODEL-FAMILY GAPS

To assess these effects in an MC pipeline we compare the two scales that compete in any comparative evaluation (`results/asm_menu_stakes.json`; real $K{=}12$ draws, the same count poll, the same generator): the menu-supplement shift of the flag's own stimulus (completing the menu with the single strongest excluded candidate) versus the accuracy gap between established model families scored on the *same* complete menus. The supplement shift is 26.8pp (ts v3, pooled answer .918 → .650), 33.6pp (wt v2, .949 → .614) and 44.0pp (CSQA, .650 → .210); Qwen2.5-1.5B-Instruct and Qwen3-0.6B-Base differ by only 3.3pp on the identical ts v3 menus. *The ordering survives several model pairs.* Extending the same-menu comparison to three families (`results/asm_menu_stakes_v2.json`; one-forward logit argmax, the paper's served answer, on the same 1500-item ts v3 menus), the largest pairwise inter-model gap is 4.0pp (SmolLM2-1.7B vs. Qwen3-0.6B), with 3.3pp (Qwen2.5 vs. Qwen3) and 0.7pp (Qwen2.5 vs. SmolLM2); and the menu-supplement shift is measured for a second complete family — SmolLM2-1.7B's own observed-to-re-menued pooled-answer drop is 21.8pp (real $K{=}12$ draws; .931 → .713), same scale as g1's 26.8pp. *Established-suite leg.* The same cell on the *established* CSQA suite (`results/asm_csqa_stakes_v2.json`; full-941 identity-safe single-token menus, one-forward length-normalized LL-argmax as the paper's served answer, all three families on the identical menus) measures each family's *own* menu-completeness shift draw-free (serve its own strongest excluded candidate, drop-weakest R0) and the pairwise gaps on the same complete menus: the shifts are 44.4pp (g1; real $K{=}12$ count@12 cross-check 44.2pp), 41.9pp (g2) and 41.8pp (SMOLLM2), while the largest pairwise gap on the identical 941 CSQA menus is 20.3pp (g1 vs. g2; 18.0pp g1 vs. SMOLLM2; 2.3pp g2 vs. SMOLLM2). *No reordering.* An evaluation-framing question is whether menu completeness "reorders or materially narrows model rankings": on CSQA the family ranking under each family's *own* re-menu is stable (g1 > SMOLLM2 > g2 on both the observed and the re-menued menus), and the pairwise gaps only close slightly under the re-menu (20.3 → 17.7pp, 18.0 → 15.3pp) — the menu-completeness shift is larger than the model gap (ratio $\approx 2.2\times$) but does not reorder these pairs on the established suite: at the *item* level, on every one of the 941 items the family gold-relation (correct vs. wrong) is identical between the observed menu and each family's own re-menu for all three pairs (0 pair-relation reversals). The under-inclusion stimulus is therefore 5–8× the largest measured inter-model gap on the *identical* near-tie synthetic ts v3 menus (26.8/4.0pp = 6.7× for g1, 21.8/4.0pp = 5.5× for SMOLLM2; no cross-suite mixing of the CSQA shift against a ts v3 gap), and $\approx 2\times$ the largest gap on the harder established CSQA suite — the ordering (menu completeness > model-family difference) survives out of corpus, with a narrower margin where families differ more. The same holds for model pairs that include the non-Qwen family: on these menus a cross-model comparison is strongly affected by menu completeness as well as by model quality, and the flag gives a gold-free indication of how much of each score depends on the menu.

Second established leg (ARC-Challenge). The same three-family cell runs on ARC ($M{=}4$, $N{=}1460$, identical length-normalized menus across the three families; `results/asm_arc_stakes_v2.json` and `results/asm_stakes_shiftlaw.json`): the family ranking on the observed menus is identical to CSQA (g1 > SMOLLM2 > g2), with again 0 item-level pair-relation reversals on all three pairs, but the shift-to-gap boundary differs. Own shifts are 15.7pp (g1; paired bootstrap 95% CI $[13.8, 17.6]$), 4.0pp (SMOLLM2; $[3.0, 5.0]$) and 3.3pp (g2; $[2.5, 4.2]$) against a largest pairwise gap of 11.0pp on the identical menus (g1-vs-g2; CI $[8.5, 13.6]$), with g1-vs-SMOLLM2 4.7pp and SMOLLM2-vs-g2 6.3pp. The under-inclusion stimulus therefore dominates the model-family gap only for the strongest family on ARC ($\approx 1.4\times$); for the two weaker families the gap is larger than their own shift. The margin resolves into each family's under-inclusion pressure $P(w^* \mid \text{re-menu answer})$, .72/.88/.95 on CSQA (deep: every family's shift 41.8−44.4pp exceeds the 20.3pp gap) versus .33/.11/.10 on ARC (shallow: shifts 3.3−15.7pp, only g1's exceeds the 11.0pp gap). *The boundary is not a point comparison.* A joint within-suite bootstrap (resampling the same item indices for every family, so each resample's shift and its largest pairwise gap are coherent) gives shift−max-gap: on CSQA +24.1pp (g1; 95% CI $[20.3, 27.7]$), +21.6 (g2; $[15.7, 27.1]$) and +21.5 (SMOLLM2; $[16.8, 25.9]$) — all three CIs exclude 0, $P(\text{shift} > \text{gap}) = 1.0$; on ARC

+4.7pp (g1; $[1.9, 7.3]$, P=1.0), −7.7pp (g2; $[-10.4, -5.1]$, P=0.0) and −7.0pp (SMOLLM2; $[-9.8, -4.2]$, P=0.0) — on the shallow under-inclusion suite the strongest family's shift significantly exceeds the largest gap while the two weaker families' shifts significantly fall short of it, with CIs excluding 0 in both directions (`results/asm_stakes_shiftlaw.json`). The item-level content is sharpest: in *every* family × suite cell the displaced candidate, when it takes the answer slot under the re-menu, is correct with frequency 0.000 (all 12 cells: N=941 × 3, N=1460 × 3, N=1000 × 3 and N=1021 × 3; this holds by construction, because w^* is absent from a menu that already contains the gold, and the CSQA item-level check of Appendix E.7 confirms $P(w^*$=item gold$) = .000$) — the shift is metered displacement of usually-correct in-menu answers by the generator's preferred off-menu continuation, never a missing gold, so shift $\approx P(w^* \mid$ re-menu answer$) \times$ acc (e.g. ARC g1 .333 × .469=15.6pp vs. 15.7); across the 12 family × suite cells (four suites, three families) $\mathrm{corr}(\text{shift}, P(w^*) \cdot \text{acc}) = 0.998$ (MAE 0.85pp), and regressing shift/acc on $P(w^*)$ gives slope 0.97 with intercept ≈ 0 pooled over the 12 cells (per-suite slopes 1.11/1.06/1.0/0.75 on CSQA/ARC/OpenBookQA/RiddleSense) — the shifted answer being never-correct is the mechanism, so the shift is the metered displacement rate $P(w^*)$ weighted by the accuracy of what it displaces. The subtlety surfaces here as the reason re-menued *means* are not a cross-family ranking: on ARC the aggregate re-menued accuracies (g1 .312, SMOLLM2 .382, g2 .326) do *not* keep the observed ordering, because the families' displacement rates differ (g1's deeper under-inclusion penalizes it most), yet no item's gold-relation reverses — the shift can move aggregate scores without reordering which family is correct on any item. Scope: on ARC no real K=12 draws exist (the count-poll cross-check is g1/CSQA only), and the boundary statement is measured on these M=4/M=5 single-suite menus.

Third established leg (OpenBookQA). OpenBookQA (Mihaylov et al., 2018) (M=4, human-written answer choices, validation + test N=1000, both held-out) is an intermediate under-inclusion suite: per-family $P(w^* \mid$ re-menu answer$) = .72/.44/.47$-equiv. (.755/.436/.468, g1/SMOLLM2/g2), between CSQA (.72/.88/.95) and ARC (.33/.11/.10) (`results/asm_obqa_stakes_v2.json`). Own shifts are 32.2pp (g1; bootstrap CI $[29.4, 35.0]$), 15.5pp (SMOLLM2; $[13.6, 17.4]$) and 16.3pp (g2; $[14.1, 18.7]$) against a largest pairwise gap of 8.7pp on the identical 1000 menus (g1-vs-g2; g1-vs-SMOLLM2 6.4pp, SMOLLM2-vs-g2 2.3pp); the joint within-suite bootstrap gives shift − max-gap = +23.5pp (g1; CI $[20.4, 26.5]$, P=1.0), +7.6pp (g2; $[3.5, 11.7]$, P=1.0) and +6.8pp (SMOLLM2; $[3.1, 10.5]$, P=1.0) — *all three* families' shifts exceed the model-family gap on the intermediate suite, so the boundary resolves into under-inclusion pressure: in the deep (CSQA) and intermediate (OpenBookQA) regimes every family's shift exceeds the largest family gap, and in the shallow regime (ARC) only the strongest family's does. The family ranking (g1 > SMOLLM2 > g2), the 0 item-level pair-relation reversals, and the collapse identity (below, 0 identity violations on all twelve cells) replicate on the third suite. The ARC "aggregate re-rank without item reordering" is *not* an ARC quirk: on OpenBookQA the aggregate re-menued means also re-rank (SMOLLM2 .205 > g2 .174 > g1 .102), because g1's deep displacement rate (.755) outweighs its accuracy lead (shift gap 32.2−15.5=16.7pp > observed gap 6.4pp), while SMOLLM2-vs-g2 keeps its order (shift gap .8 < 2.3pp); on CSQA the shift gaps are all ≤ 2.6pp and no pair re-ranks. The same coexistence appears on RiddleSense (SMOLLM2 re-menued .146 > g1 .1205): three of four suites show the coexistence, one does not, and the decomposition's flip condition decides each pair.

Closed-form boundary. Since $\text{shift}_f \approx P(w_f^* \mid \text{re-menu answer}) \cdot o\text{-acc}_f$, the per-family boundary "the menu-completeness shift exceeds the largest family gap" is decided by

$$P(w_f^* \mid \text{re-menu answer}) \;>\; \frac{\text{max-gap}}{o\text{-acc}_f},$$

which matches the joint-bootstrap verdict in *all* twelve cells (CSQA/OpenBookQA: every family satisfies it; ARC: only g1, $.333 > \frac{.11}{.469}$ vs. SMOLLM2 $.106 < \frac{.11}{.422}$ and g2 $.103 < \frac{.11}{.359}$; RiddleSense: only g1 $0.8198 > \frac{20.8}{.483}$ since the weaker families' $P(w^*)$ falls below gap/acc) – a one-line rule for when the re-menu audit moves a checkpoint's score by more than the largest model gap on the observed menu.

F.6 Abstention utility and its benchmark-dependent limits

The ratio threshold tests whether the strongest examined excluded candidate outranks the menu leader. The implemented rule flags $\log \tau > 0$ and retains ties. Its original-answer accuracy is compared with several selectors in Appendix F.10. This experiment uses full-answer CommonsenseQA

validation scores (1197 items before the seeded evaluation split), not the 941-item single-token intervention slice. The evaluation pool contains both single- and multi-token answers.

The recorded excluded-score comparison selects the lowest maximum off-menu log score, not the highest within-menu margin. On the original evaluation pools, ratio-clean accuracy is .684 versus .561 on CommonsenseQA, 0.4654 versus 0.4604 on ARC, 0.453 versus 0.4274 on OpenBookQA, and 0.68 versus 0.45 on RiddleSense, at matched coverage $35.4/68.2/23.4/19.6\%$. These differences are $+12.3$, $+0.5$, $+2.6$, and $+23.0$ percentage points. They describe a weak excluded-score baseline. The within-menu margin instead outperforms the ratio on the first three suites and is unresolved on RiddleSense (Figure 3). Low excluded score alone conflates leader strength with excluded-candidate strength; dividing by the leader changes the ranking, but that fact does not establish a superior correctness selector.

An excluded-score cutoff transferred from CommonsenseQA gives coverage .538 on ARC and .332 on OpenBookQA, versus ratio-clean coverages .682 and .234. This illustrates score-scale transfer limits. The common unit ratio threshold has an interpretable candidate-ordering meaning, but neither its coverage nor its error rate is constant across tasks.

Random abstention is a weaker reference. A 2000-subset Monte Carlo comparison finds an advantage on CommonsenseQA and RiddleSense, but not on ARC or OpenBookQA. Appendix F.10 gives the exact hypergeometric counterpart and the stronger margin comparison. These conditional random-subset intervals do not estimate population uncertainty. The normal-approximation minimum detectable effects for served counts 795 (ARC) and 117 (OpenBookQA) are 5.0 and 12.9 points at nominal two-sided 5% level and 80% power. They describe sensitivity under a binomial approximation, not equivalence bounds. Task and answer-length differences are confounded in these data, so the results cannot be attributed to a single-token versus multi-token distinction.

F.7 Serving with multi-token answer scores

The same serving rules run on the full multi-token artifacts (identity-self-contained: no cross-run join; `results/asm_fullserve_multi.json`, `results/asm_serve_deferral_obqa.json`). The baseline serves the *highest*-margin items. Serve by the length-normalized likelihood margin with the $\tau \geq 1$ flag, a flag-deflated gate, and deferral to the clean pool (F1), the Section 6 policies; a single per-call read makes any flag-conditional deflation collapse to the margin gate ($F \equiv B$, bootstrap difference 0), so the flag's only serving action at multi-token scale is deferral. *The result is a boundary, not a transfer.* On the two natural multi-token suites with fully independent labeled test splits (ARC: val→test positional; OpenBookQA: val→test), deferral moves matched-coverage served accuracy by roughly -6 to $+5$pp depending on family and suite: ARC test at 50% coverage $B{\to}F1 = .529 \to .508$ (g1), $.387 \to .403$ (g2), $.455 \to .475$ (SmolLM2); OpenBookQA test $= .512 \to .453$ (g1), $.395 \to .379$ (g2), $.467 \to .519$ (SmolLM2). The mechanism is that on these suites the clean pool is *not* more accurate than the flagged pool (whole-suite clean-minus-flagged LL-accuracy: ARC g1 -0.3pp, OpenBookQA g1 -1.0pp), so deferral only re-ranks within the margin frontier, where the host checkpoint's flagged high-margin items are served well by the margin (.519–.745 within-window). On the recorded CSQA basis the clean pool is more accurate ($+12.1$pp for g1) and deferral there is the measured abstention action of Section E.7; CSQA's public test answers are not available, so its full suite is evaluated on a seeded half (CSQA-full $B{\to}F1$ at matched $50\% = .766 \to .788$). The deferral results do not establish an original-answer correctness gain beyond the within-menu margin. The flag diagnoses a different outcome, response sensitivity to candidate replacement. Neither the observed positive comparison with random abstention nor the coexistence of single- and multi-token answers identifies answer length as its cause. Comparing sequence-score and first-token diagnostics on matched multi-token items remains open.

F.8 Qwen3-4B on established benchmarks

The candidate-omission sensitivity also appears on Qwen3-4B (the same model family as the smaller Qwen3-0.6B checkpoint), on the identical *established* menus (`results/asm_obqa_stakes_q4b.json`, `results/asm_csqa_stakes_q4b.json`; same length-normalized LL-basis runner, no tuning). On OpenBookQA the stronger family ranks **q4b** $>$ g1 $>$ smol $>$ g2 on the observed menus (acc 0.441 vs. .424/.360/.337) yet its *own* menu-completeness shift is still 26.1pp at $P(w^* \mid \text{re-menu answer}) = 0.562$ — above the enlarged 6-pair maximum gap of 10.4pp, so the

intermediate-regime boundary holds for the modern model as well; on CSQA the shift is 46.8pp (acc 0.6536, $P(w^*) = 0.7736$), above the 20.3pp gap. Item-level pair-relation reversals remain 0 on all 6 families-pairs per suite (the family gold-relation is unchanged by each family's own re-menu even across a 4B model), and the aggregate re-menued means re-rank on OBQA (smol .205 > q4b .180 > g2 .174 > g1 .102). Scope: Qwen3-4B is a post-trained reasoning checkpoint in the Qwen3 line (the other Qwen legs here are much smaller, g2 is a base model); the reproduction targets the menu-completeness *stimulus*. *ARC leg and the shallow-regime boundary.* On the shallow ARC suite the modern family is the *least* under-included ($P(\tau \geq 1) = .17$, $P(w^* \mid \text{re-menu answer}) = 0.1705$, acc 0.5055 — it ranks above g1 on the observed menus with 0 pair-relation reversals on all six pairs), and its own shift 8.9pp is *below* the pairwise gap, which the stronger family itself enlarges to 14.7pp (re-menu gaps shrink to $\leq$ 10.4pp). Only the deepest-under-included original family (g1) still exceeds the gap (+1.0pp) — the boundary is decided by under-inclusion *pressure*, not model scale (`results/asm_arc_stakes_q4b.json`). *RiddleSense leg (fourth suite).* On the single-token riddle suite the modern family ranks second on the observed menus (acc 0.4437 vs. g1 .483/smol .331/g2 .275, with 0 reversals on all six pairs), its under-inclusion is intermediate ($P(w^* \mid \text{re-menu answer}) = 0.572$) and its own shift is 22.4pp — numerically above the *unchanged* 20.8pp gap (+1.7pp; paired bootstrap 95% CI $[-2.2, 5.7]$pp, P=.81), i.e. marginal, not significantly above. The modern family does not enlarge the riddle family gap the way it does on OpenBookQA/ARC — at second-place accuracy the widest family pair remains g1-vs-g2 — so the boundary statement keeps the same form: the deepest-under-included, most able family significantly dominates; everyone else is within uncertainty of the gap (`results/asm_riddle_stakes_q4b.json`). *Abstention-utility transfer.* The unit-threshold abstention cell (Appendix F.6) reproduces on the modern family (`results/asm_tau_utility_q4b.json`): on CSQA the ratio-clean pool serves at 0.822 (coverage 0.2505) vs. the matched-size random pool 0.6764 (95% CI $[.602, .746]$) and the matched-coverage mass oracle 0.6864 — +14.6pp over random, outside its CI — and on OpenBookQA it is again within noise of random/always (0.426 vs. 0.4573/.458; matched mass oracle 0.4484). The same neutrality holds on ARC's test slice (n_{srv}=969, flag rate 16.8%): served 0.509, paired delta −0.5pp, 95% CI $[-1.8, 0.8]$pp, P=0.7795 — same sign pattern as the host family's ARC cell (0.0pp, $[-1.9, 2.0]$) (`results/asm_tau_utility_q4b_arc_delta.json`). On RiddleSense the single-token selection effect transfers to the modern family (`results/asm_tau_abstention_riddle.json`): the ratio-clean pool serves 0.5357 vs. a matched-size random pool 0.4706 (paired delta +6.5pp, 95% CI $[1.8, 11.6]$pp, P=0.005) and a matched-coverage mass oracle, same sign as the host family's cell (0.68 vs. 0.5071, +17.3pp, $[8.0, 26.0]$pp). The home-suite advantage and the multi-token neutrality (both OpenBookQA and ARC) of the abstention layer therefore survive a 4B modern reproduction of the same sign and approximate magnitude as the 1.5B host family. *Scoping negative (weakest family).* The deepest-under-inclusion family breaks the transfer: on the identical RiddleSense items the SmolLM2-1.7B ratio-clean pool (0.444=44pp coverage; flag rate 55.6%) serves 0.366 vs. a matched-size random pool — paired delta +3.3pp, 95% CI $[-1.3, 7.9]$pp, P=0.098 — within noise, against the host (+17.3pp, P<.001) and modern (+6.5pp, P=0.005) cells; on the SmolLM2 CSQA subset the unit flag fires on 87.7% of items, leaving a 0.119-coverage clean pool that serves 0.464 vs. random (delta −5.4pp, CI $[-17.9, 7.1]$pp, P=0.834) (`results/asm_tau_utility_smol_delta.json`). The base family adds a third profile: Qwen3-0.6B-Base (g_2) shows a large-but-fragile CSQA gain — clean pool 0.731 at 26 items (94.7% flagged), delta +27.0pp, CI $[7.7, 46.2]$pp, P=0.0045 — against a null Riddle cell (delta −1.5pp, CI $[-8.0, 4.4]$pp, P=0.72) (`results/asm_tau_utility_g2_delta.json`). The selection outcome is therefore *family/profile-specific*, not a free property of the ratio: it holds for the two moderately under-included instruct families (g_1, Qwen3-4B), is absent for the deepest-under-inclusion instruct SmolLM2, and is fragile and split for the base family g_2 (Table 18).

F.9 Few-shot format robustness of the displacement law

A related question is whether the off-menu displacement law is an artifact of the paper's one-forward "Question–Answer" reading. We rerun the same g_1 runner on a deterministic 95-item subset of CSQA whose prompts carry a fixed, *format-only* set of five hand-written multiple-choice demonstrations — no benchmark training rows, so the change is pure format, not task exposure (`data/asm_fewshot_csqa5_gen100.jsonl`, `results/raw/asm_full_csqa941_ll_g1_fewshot5.json`): on the identical items the one-shot (format A) and generic-5-shot (format B) reads agree on the flag decision for 0.8105 of items, the flag rate and under-inclusion depth

family	profile	CSQA-941 Δpp [CI], P	RiddleSense Δpp [CI], P
g_1 Qwen2.5-1.5B	moderate	$+6.9$ $[1.4, 12.3]$, P=0.009	$+17.3$ $[8.0, 26.0]$, P<.001
Qwen3-4B (q4b)	moderate	$+14.8$ $[7.6, 22.0]$, P<.001	$+6.5$ $[1.8, 11.6]$, P=0.005
SmolLM2-1.7B	deepest	-5.4 $[-17.9, 7.1]$, P=0.834	$+3.3$ $[-1.3, 7.9]$, P=0.098
g_2 Qwen3-0.6B	base	$+27.0$ $[7.7, 46.2]$, P=0.0045 (n=26)	-1.5 $[-8.0, 4.4]$, P=0.72

Table 18: Abstention profile map on the two single-token suites: paired clean-pool served-accuracy deltas vs. matched-coverage random pools (95% random-subset intervals and $P(\Delta \leq 0)$, 2000 seeded draws; ARC/OBQA show no detected advantage, with sensitivity estimates in Appendix F.6). Sources: `results/asm_tau_utility_delta.json` (g_1), `results/asm_tau_utility_q4b_csqa_delta.json`, `results/asm_tau_utility_q4b_arc_delta.json`, and `results/asm_tau_abstention_riddle.json` (Qwen3-4B), `results/asm_tau_utility_smol_delta.json` (SmolLM2), `results/asm_tau_utility_g2_delta.json` (g_2).

$P(w^* \mid$ re-menu answer$)$ are if anything *higher* under the few-shot format ($0.7263 \rightarrow 0.7895$), and the own menu-completeness shift is *identical* (35.8=35.8pp). The equality is exact at the stored precision, and structural rather than coincidental: the format moves the served and the re-menued accuracy together — $0.5895-0.2316$=$0.5474-0.1895$=$.3579$ (equivalently, both accuracies fall by exactly 4.21pp in lockstep) — so the displacement magnitude is invariant even though the served level moves. Served accuracy moves by -4.2pp (3 items better under few-shot vs. 7 under one-shot among the 10 changed items; conditional sign-test P=0.3438), i.e. within noise: the mechanism is format-robust on the measured subset, and the format changes the served answer's absolute level, not the displacement structure the flag and shift describe. Testing the three readout moves: the flag-rate (and depth) uptick rests on 12 vs. 6 discordant items among 95 (exact two-sided McNemar P=0.2379); the re-menu move has net -4 items, but over every McNemar table with that net the exact P ranges $[0.1250, 0.7572]$; the accuracy move is the sign split tested above (P=0.3438). Power is computed, not assumed: holding the observed discordant count fixed, a flag-rate net of $+10.5$pp or larger (14:4 discordant items) would already clear significance (P=0.0309); the measured net of $+6.3$pp sits well inside the unresolved band. None of the three moves is detectable — on 95 items a shift of this size is not resolvable, so the read is "no detectable format effect on any readout," not "format invariance". Scope: 95 items, one generator, generic (not benchmark-native) exemplars; other prompt formats are not tested.

F.10 Equal-coverage controls for original-answer accuracy

This comparison uses the original evaluation pools and mean-token-log-likelihood scores. The ratio retains $\tau \leq 1$; all alternatives retain the same count. They differ only in score ranking. The within-menu margin is the top-two difference after softmax over listed mean log likelihoods. The raw gap omits that softmax, and confidence is the largest softmax value. The excluded-score rule retains the lowest maximum excluded score. Cutoffs use evaluation scores without labels and therefore describe these fixed pools rather than a separately calibrated deployment policy. Table 2 in Section 5 reports original-answer accuracy for all five selectors at identical served count within each suite (served/evaluation counts 212/599, 795/1165, 117/500, and 100/511), using one fixed host checkpoint rather than multiple training seeds. The 5000 paired item bootstrap replicates resample evaluation item indices and retain the original selector masks. Each replicate evaluates both selected pools on that common resample, using its realized selected denominators. Intervals are conditional on the fixed checkpoints and selectors, which require no fitting; cutoffs are not reselected in each bootstrap. The random reference instead samples a fixed-size subset of the finite evaluation pool, giving an exact hypergeometric distribution. Its interval describes random-subset variation and is not a confidence interval for a population accuracy.

The same table gives the random-subset reference at the same served counts: the hypergeometric mean, its central 95% range, and the upper-tail probability of the ratio-clean accuracy, conditional on each observed evaluation pool (unrounded tail probabilities 0.00593, 0.51846, 0.59043, and 0.00008 for CSQA, ARC, OpenBookQA, and RiddleSense).

G SEMANTIC VALIDITY OF EXCLUDED CANDIDATES

G.1 SEMANTIC-VALIDITY JUDGMENTS AND JUDGE CALIBRATION LIMITS

The stakes result of Appendix F.5 decomposes the menu-completeness shift through w^* — each family's strongest excluded candidate, which by construction is never the item's gold (it is absent from a menu that contains the gold; an item-level check confirms $P(w^*{=}\text{item gold}) = .000$ on all twelve family/suite cells). The open question is whether the menu *omits a legitimate competing answer* (w^* semantically answers the question) or the generator only prefers an irrelevant continuation. *Protocol 1 (single-candidate yes/no) fails.* A blind first-token —"is this a reasonable and valid answer?"— read from two independent instruction-tuned families (SMOLLM2-1.7B-INSTRUCT and LLAMA-3.1-8B-INSTRUCT) does not separate the gold-answer probes from presented non-gold options: judge recall = false-positive on every suite for both (recall=fp=.51/.53/.20 and .98/.99/.94, gold-vs-distractor AUC = .49 both) — the yes/no protocol is governed by an acquiescence "coherent candidate" prior, its raw "valid" rates are uninterpretable, and a recall-adjusted rate is undefined. *Sensitivity of the yes/no audit.* The observed AUC is near chance (.49). The probe-budget calculation in `results/asm_mde.json` gives an approximate 80%-power direction-accuracy threshold of 58.1% under a binomial model. This sensitivity calculation neither bounds the true discrimination at chance nor supplies an AUC confidence interval; paired probe dependence must be retained in any such inference. *Protocol 2 (forced pairwise,* LLAMA-3.1-8B-INSTRUCT*) cancels the prior.* The judge must *choose*: "which candidate better answers the question, A or B", read as the first-token A-minus-B logit. On the probe pairs the judge separates gold from distractor with 76–83% direction accuracy (gold chosen on 67.0%/67.0%/66.0%/60.0% vs. distractor chosen on 1.3%/5.3%/2.8%/8.2% of the probe pairs; weakest on the single-token riddle probes, whose human distractors are themselves plausible answer words), so the pairwise read is a weak but informative validity ranker. On the flagged items the judge ranks the item's *gold* strictly over w^* on 80.5%/86.6%/67.3%/79.1% of items and w^* over or tied with the gold on the remaining 19.5%/13.4%/32.7%/20.9% (95% CI $[16.7, 22.4]\%/[10.5, 16.5]\%/[29.3, 36.0]\%/[18.2, 23.7]\%$), with a mean A-minus-B logit of $-2.26/-3.67/-.89/-1.70$ nats — i.e. a *sizable minority* of the flagged omissions (one in five on CSQA and on RiddleSense, one in seven on ARC, one in three on OpenBookQA) are judged gold-equivalent-or-better answers the closed menu never contains, while the majority are strictly below the gold. The verdicts are not threshold noise: the $w^* \geq$ gold items carry median A-minus-B margins of $+1.48/+1.36/+2.22/+1.11$ nats — spread into the positive direction — narrow near-tie bands ($|m|<.5$ nats) account for only 8.7%/6.4%/11.5%/12.8% of flags, and the complementary majority is strongly gold-favouring (47.3%/67.3%/35.8%/36.6% of flags below -2 nats; `results/asm_offval_margins.json`). The *error-correction significance* stays a measured null on either reading: whether w^* is a valid-but-weaker candidate or a gold-equivalent alternative, in the measured cells it is never the item's benchmark gold, so on those cells serving it never repairs a wrong in-menu answer — the under-inclusion cost is the wholesale displacement of usually-correct in-menu answers (accuracy hostage fraction shift/acc, Appendix F.5), not a missing-correct-answer failure. The flag's residual role is the menu-dependence audit, not answer repair. *Cross-suite ordering.* Across all four suites the gold-equivalence rate orders 13.4% < 19.5%≈20.9% < 32.7% (ARC and OpenBookQA are separated from the central pair by non-overlapping 95% CIs; CSQA and RiddleSense overlap near one-in-five), and it is *not* monotone in under-inclusion depth ($P(w^* \mid \text{re-menu}) = 0.3329/0.7205/0.755/0.8198$): RiddleSense has the deepest under-inclusion of any suite yet only a CSQA-level rate, because option-set plausibility modulates the read — its single-token distractors are all plausible answer words (weakest probe separation for *both* calibrating judges: 76% for the 8B and 56.2% for the 1.5B decoder, below their other three suites), so its strongest omitted candidates meet stronger competition from the gold answer. The CSQA → OpenBookQA rise at nearly equal depth is the same plausibility amplification: OpenBookQA's distractors are closer to their golds, so its strongest omitted candidates are more often judged as good as the gold. *Near-independence of the read.* The pairwise verdict is not a restatement of the other signals related to validity (`results/asm_offval_crosstab.json`): w^* being a gold-answer word elsewhere (0.332/0.656/0.558 of flags on CSQA/ARC/OpenBookQA) and the judge's $w^* \geq$ gold verdict are only weakly associated (18.2 vs. 20.1pp on CSQA, 14.4 vs. 11.4pp on ARC, 35.4 vs. 29.3pp on OpenBookQA; gaps of one to a few pp, mixed sign on the in-menu-answer-wrong split 21.5 vs. 18.2pp), and the judge margin correlates only weakly with the generator's off-logit margin ($r{=}.15/-.18/.32$) — the blind judge supplies a read of answer validity that is not redundant with the displacement statistics. *Judge*

robustness is calibration-specific. The pairwise verdicts rest on a single judge family; two further instruction-tuned decoders re-read the identical blinded probes under the same forced-pairwise protocol, and passing the probe audit turns out to be a checkpoint property rather than a scale guarantee. SMOLLM2-1.7B-INSTRUCT *fails* it on all four suites: gold-vs-distractor direction accuracy 50.3%/49.3%/49.6%/49.7% (vs. $83/81/81/76\%$ for the 8B judge), pooled 49.7% over 1178 probes, with 95% CIs that all include 50% ($[44.7, 56.0]\%/[43.7, 55.0]\%/[43.9, 55.4]\%/[43.9, 55.4]\%$). The failure mode is a first-position heuristic, not noise: on the CSQA probes the decoder chooses position A with probability 76.0% whether A holds the gold or a distractor ($P(A|\text{gold})$=76.0% vs. $P(A|\text{dist})$=75.3% for each question's gold/ distractor probe pair), so its choices carry no content information and its $w^* \geq$gold verdicts on the flagged items would be uninterpretable. The margin carries no residual content either: the decoder's confidence is monotone in its position prior — on the highest- |A−B| probe tercile it picks position A on 100% of both the gold-in-A and the distractor-in-A items ($P(A|\text{gold})=P(A|\text{dist})=1.0$, direction accuracy still .50), so no margin-thresholding recovers a validity signal (`results/asm_smol_probe_margin_analysis.json`). Within this paper's probe budgets, this decoder shows no measured pairwise separation. The approximate direction-accuracy sensitivity threshold is 58.1%; it is not a bound proving chance-level population discrimination. A second small decoder of an independent family, QWEN2.5-1.5B-INSTRUCT (the g_1 generator), *does* separate under the identical protocol: direction accuracy 57.3%/61.3%/60.8%/56.2% (pooled 58.9%), above chance on every suite (two-sided binomial; CSQA P=0.011, riddle P=0.035, ARC/OpenBookQA $P<.001$), with content- vs- position gaps $P(A|\text{gold})-P(A|\text{dist})=14.7/22.7/21.7/12.3$*pp* and weakest on the single-token riddle probes, as for the 8B judge (CSQA gap $67.0-1.3=65.7$pp) (`results/asm_g1_probe_gate.json`). The pairwise protocol is thus open at modest scale: SmolLM2's failure is its own positional prior, not a scale limitation. *Self-interest ceiling.* The generator's own 1.5B decoder re-reads the same flagged items (`results/asm_g1_wstar_judge.json`); its $w^* \geq$gold rates are $45.0\%/30.0\%/35.7\%/43.0\%$ (pooled 38.9%), systematically above the independent 8B judge's $19.5\%/13.4\%/32.7\%/20.9\%$, up to $\approx 2.3\times$, combining its self-interested preference for the candidates its own generation excludes with its positional prior. Consistent with a displacement-driven read, the generator's self-certification rate is *rank-ordered by its own shift* across the four suites ($44.4/15.7/32.2/36.2$pp $\leftrightarrow$ $45.0\%/30.0\%/35.7\%/43.0\%$, rank agreement ρ=1.0 on n=4, one-sided permutation p=.04), whereas $P(w^*|\text{re-menu})$ alone is not (ρ=.4): the ceiling tracks displacement magnitude, not raw under-inclusion. The closed-set- incompleteness reading therefore survives at two judge scales: the 8B rate is the conservative end ("up to one in three"; its OpenBookQA peak 32.7% is matched by the generator's own 35.7%), the generator's own ≈two-fifths pooled rate is a self-interest ceiling, and the *strict* two-judge gold-equivalence consensus is 11.4% of flagged omissions (A/B agreement 61.9%); either way, no judge ever finds w^*=the item's gold.

H INPUTS AND IMPLEMENTATION DETAILS

The experiments use the forced-choice item files, cumulative $K = 12$ draw-count paths, per-item extended-set scores, and controlled re-menu mappings included in the supplementary material (a zip archive). In that archive, raw model outputs are under `results/raw/`, derived numerical results are under `results/`, constructed item files are under `data/`, and the analysis implementations are under `scripts/`. The filenames cited beside individual comparisons in this appendix refer to these files and identify their exact inputs. The README of the supplementary material maps every table and figure to its script and result file, and `scripts/reproduce_check.py` lists the command that produces each result file.

Decision statistics and labelled fitting. The raw menu scores, mass margin, and common-normalizer off-menu threat use no evaluation labels. Gold labels define accuracy, the analysis-only gold-relative threat, and the intervention's retained gold option. Calibration and confidence deflation use a separate labelled fit split. The equal-coverage original-answer controls instead choose cutoffs from evaluation scores without evaluation labels; those comparisons are conditional on the resulting fixed pools. Each serving result states its coverage, comparator, and uncertainty basis; evaluation-item bootstrap intervals conditional on a fit do not include calibration-fit uncertainty.

Exact scores and sampling. The single-token controlled menus use masked probabilities from a common next-token forward. Multi-token comparisons use the specified sequence-score convention

and are not interchangeable with first-token masked scores. Candidate search and answer scoring are part of the measured protocol; a flag is always relative to the examined candidate set. Stored count paths permit analysis of smaller draw budgets without new model calls. Re-menu probabilities are renormalizations of the same per-item forward when menu text is absent from the prefix; agreement between predicted and sampled responses checks that protocol rather than providing an independent causal result.

I EXTENDED DISCUSSION OF LIMITATIONS

This section gives the scope conditions behind the future-work directions of Section 8.1.

Scope and model access. The results require candidate-score access and cover four small open-weight checkpoints (0.6B to 4B) from three families. Transfer to larger or API-only models and independent training replications remains untested. Of the 941 single-token CommonsenseQA items, 838 are training rows, and contamination cannot be excluded for any benchmark; on the 103 validation rows alone the host's pooled-answer drop is 48.5 points (43.9 on training rows; Appendix E.7).

Excluded-candidate search and scoring. The excluded set is finite and filtered; neither flag convention ($\tau > 1$, $\tau \geq 1$) certifies completeness. τ compares the strongest examined excluded candidate, not the total probability of excluded answers. For multi-token answers, $q_{\mathcal{O}}$ and τ compare mean token log likelihoods, not complete-answer probabilities. Normalization, answer length, and candidate search can affect which excluded candidate wins; they change together with task, candidate construction, and tokenization (full-answer CommonsenseQA mixes single- and multi-token golds) and are not isolated, so neither an answer-length explanation nor transfer to open-ended semantic uncertainty is established.

Menu intervention and semantic validity. The intervention inserts the model's top excluded candidate, which is never gold. Its effect is thus partly built into the design; the finding is the shift's size, not its direction. Large effects are task- and checkpoint-dependent rather than inevitable: on ARC, three checkpoints lose less than the 14.7-point original-menu gap. Checkpoint-specific altered menus do not define a common-stimulus model ranking, and prediction–draw agreement, which restricts the same forward scores, does not test an independent causal mechanism. An excluded candidate can change the response while remaining incorrect, and the ratio does not establish that any excluded candidate is a better answer. Semantic judgments are model assessments from one 8B judge (76–83% probe accuracy), not human annotations; a SmolLM2 judge is at chance, and agreement with the host as a second judge lowers the pooled rate to 11.4%.

Selection and calibration. The ratio is not a better correctness selector than the within-menu margin (three of four suites; RiddleSense is unresolved, $[-12.4, 10.4]$); its gain over random subsets is not detectable on ARC or OpenBookQA and does not outweigh the margin control; and the unit threshold's coverage and error rate vary across tasks. The equal-coverage comparison uses fixed pools, not a separately calibrated deployment threshold. Re-menu calibration transfers unevenly: the Qwen3-4B interval includes zero, and Qwen3-0.6B and SmolLM2 flag nearly every item, so their clean pool supplies only about 5% coverage, deflation has little to reorder at 50% coverage, and draw risk rises slightly (.903 to .914; .883 to .893); sign and size depend on checkpoint and coverage, since a small unflagged pool cannot supply a large low-risk served set. On the cloze tasks, larger reductions cost original-menu accuracy, and the fitted 5% risk targets fail to attain 5% on the evaluation halves; these operating points are empirical predictors, not distribution-free guarantees.

Exact scores and sampling. Exact argmax removes sampling noise, not model error, and can be less accurate against gold: without TinyStories v3 the pooled cloze interval widens to $[-0.11, 0.74]$ and SmolLM2 reverses the direction on single-token CommonsenseQA, so no universal gold-accuracy advantage is established, and the cost of fixed-menu draws does not establish superiority over open-ended sampling methods.

Statistical scope. Intervals condition on the fixed pools, checkpoints, and protocols. Random-subset intervals and Figure 4 shading (the min–max over three partitions) are not population con-

fidence intervals. Repeated splits and draw seeds do not measure training-seed uncertainty; menu-sensitivity results describe fixed datasets.