BENCHMARK VISIBILITY FROM ANNOTATIONS ALONE: AN AUDITABLE LOWER BOUND ON BLIND EXPLOITABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Whether a benchmark can *see* the capability it is cited for is a property of the corpus, and the field establishes it by running systems that ought to fail: late, expensive, and one corpus at a time. We show that much of the same information is computable from the annotations alone: no model, no judge, no frames, which matters because every corpus here keeps its media behind a signed agreement and its annotations public. The instrument is two counts and a curve. *Announcement* counts items whose gold answer is stated in the question; the *blind bound* is the best accuracy of an enumerated family of strategies that see the question's form and the answer distribution and nothing else. It is one-sided (it condemns and can never acquit, for a reason no instrument of this kind escapes), and announcement is a detector of the same failure rather than independent evidence, so the two counts must not be added, yet they flag nearly disjoint items. The curve is what the field is missing: the same estimator returns a positive value when the template carries no information at all, a null-induced baseline fixed by the group-size profile alone, and it collapses as the template isolates items; we construct a corpus on which it rises first instead, and which shape a given corpus takes is measured rather than deduced. Either way the same corpus reads very differently under two choices papers rarely state. Three common estimators mistake a corpus's answers for its structure by one mechanism: in-sample scoring and the ordinary nonparametric bootstrap leak the held-out answer back into its own evidence, and leave-one-out inside a group of two fits the single other item. The bootstrap is the dangerous one, because it inflates the bound substantially while looking like rigour, and deduplicating its own draws returns it almost exactly to the point estimate; a fit-and-test split is the repair. We validate the instrument against a documented defect on two corpus versions identical outside the repair, against a third-party repair its own authors performed, and behaviourally on an open-answer corpus, where the separation survives an intervention that moves the surface feature the mask is built from while leaving the question's content alone. Across the corpora we survey, 3 of the 5 option-regime corpora carry a statistically supported blind-exploitability signal under a family-wise max-statistic permutation test that prices the rule family's own selection into the null.

1 INTRODUCTION

A benchmark is cited for a capability. Whether it can *see* that capability is a property of the benchmark, not of the systems evaluated on it, and the field's usual way of finding out is to run a system that should fail and observe that it does not. That method works, and has worked repeatedly (Anand et al., 2018; Majumdar et al., 2024; Udandarao et al., 2025; Lee, 2026; Xu et al., 2026), but it has three costs. It is expensive: the most careful recent instance runs twenty models across four viewing conditions (Lee, 2026). It is late: the finding arrives after the corpus has been published, adopted and cited, and often after a generation of systems has been tuned against it. And it is local: each result concerns one corpus and one baseline, so the next corpus starts over.

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.

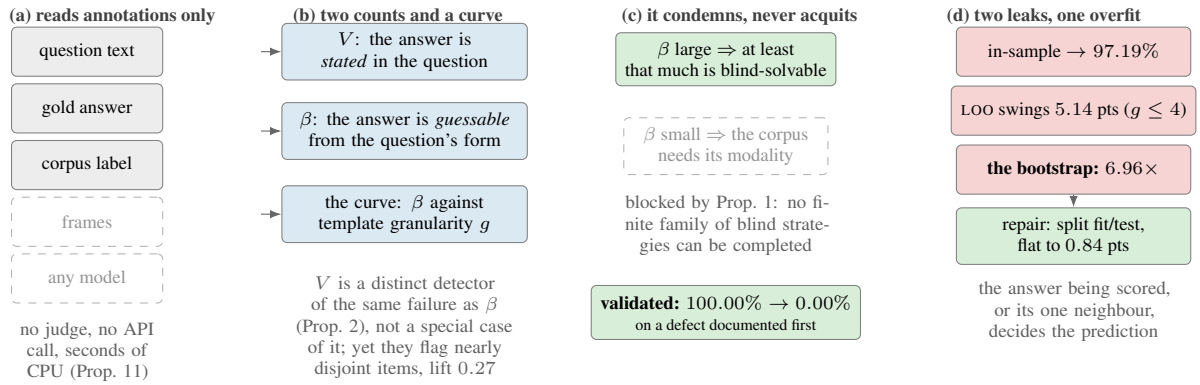

Figure 1: The instrument, and what it costs to get it wrong. **(a)** It reads a corpus's annotations and nothing else (no frames, no model, no judge), which is what lets it run on corpora whose media sit behind a signed agreement, and lets a corpus's authors run it before release. **(b)** It reports two counts and a curve: how often the answer is stated in the question, how often it is guessable from the question's form, and how the second varies with the grouping used to compute it. **(c)** The inference runs one way only. A large β condemns; a small one acquits nothing, because no finite family of blind strategies can be completed. We validate the instrument against a defect that was documented before it existed, on two corpus versions that are identical outside the repair. **(d)** The result we did not expect. One mechanism (the answer being scored, or the single neighbour standing in for it, deciding the prediction) has three routes: two are leaks in the sense of Definition 3 and the middle one is an overfit. The third is the ordinary nonparametric bootstrap, which inflates β by $6.96\times$ while looking like rigour. A fit-and-test split is the repair. The leave-one-out swing is over $g \in \{0, 1, 2, 4\}$, not over the whole sweep, where it is 10.21 points.

This paper computes a large part of the same information from the annotations, before anything is run. The instrument reads questions, gold answers and whatever grouping labels the corpus supplies. It never reads a frame: not a limitation we accepted but a property we needed, since the frames of every corpus here sit behind a signed agreement and the annotations do not.

What the instrument is. Two counts and a curve. *Announcement* V counts items whose gold answer appears verbatim in the question; on those items the answer was stated, so nothing beyond reading the question is being tested. *The blind bound* β is the accuracy of the best strategy that sees the question's form and the answer distribution and nothing else: in an open-answer corpus, the modal answer within a group of items sharing a question template, estimated leave-one-out so an item never votes for its own answer; in an option corpus, the label space is the option index and the strategies are the surface rules a blind system can apply, with the bound their maximum, which is what keeps it one-sided. *The curve* is β against template granularity, read against a permutation null built from the corpus's own answers. A blind baseline is always computed at some granularity, usually implicitly, and the group-size profile alone fixes a positive value the estimator returns before any template–answer association is present.

The number is not a number until its granularity is stated. This is the paper's sharpest practical claim, and it is partly a theorem and partly a measurement. Theorem 1 proves what follows from the estimator alone: refining a template can only shrink groups, the leave-one-out estimate is positive under the permutation null before any structure is present, and in the limit each group holds one item and the estimate collapses to zero. What the theorem does *not* prove is the rise in between. Proposition 7 shows the rise is available (an eight-item corpus goes $62.5\% \to 87.5\% \to 0\%$), and Table 1 says which corpora turn at all: 5 of 9 rows are non-monotone and 4 are not. On OpenEQA, the same items at the same fine granularity yield 1.6% held out and 97.2% scored in sample, and 11.7% at the coarsest granularity: three numbers from one corpus, from two choices papers reporting blind baselines rarely state. None of them is a lie. None of them is a measurement.

Contributions. (1) **An instrument** for blind solvability requiring no model, no judge, no frames and no network, therefore computable on corpora whose pixels we cannot legally hold; it is CPU-only and linear in the permutation budget, a 47863-item, 9-row audit costing just under two hours on one

core. (2) **One-sidedness** (Proposition 1): the bound condemns and never acquits, and we say why acquittal is unavailable to any instrument of this kind rather than treating it as an implementation gap. (3) **Announcement is a distinct exposure detector alongside the blind bound** (Proposition 2), so the two counts must not be added, and no ordering of their rates holds, with the measured contingency showing the two to be nonetheless nearly disjoint. (4) **The refinement theorem and the separate possibility result** (Theorem 1, Proposition 7). (5) **Two leaks, an overfit, one mechanism, and a repair** (§4.3): in-sample scoring, leave-one-out inside a group of two, and (new, and to our knowledge unreported) the nonparametric bootstrap, whose inflation we diagnose by deduplicating its own draws; the split estimator is the repair. (6) **A positive control with independent ground truth** (§4.2), on a corpus whose defect was documented before the instrument existed. (7) **The instrument run on 8 corpora in 9 rows**, 47863 items across both answer regimes, of which 7 rows were unseen when the estimator, the null and the refinement analysis were fixed. (8) **What the bound corresponds to behaviourally, and what carries it**: the flagged items are the ones blind backbones actually solve, and because that is a correlation we intervene on it under two designs, the second dose-matched token for token.

What this does not claim. A small β is not evidence that a corpus requires its modality. That is the content of Proposition 1 and it is a limit on the instrument, not a caveat about its implementation. Nor does the instrument replace a model-based diagnostic: a visual-dependency gap measured across many models (Lee, 2026) answers a question about *models* that annotations cannot reach. The claim is narrower and, we think, useful: the part of the answer that is a property of the corpus alone should be computed before the corpus is published, and it costs nothing to compute.

2 THEORY

Definition 1 (Corpus, blind view, blind strategy)**.** A corpus is a finite set of items $x = (q, a, \ell)$ with question q, gold $a \in \mathcal{Y}$ and grouping label ℓ supplied by the corpus; the *blind view* of an item is (q, ℓ). A *blind strategy* is a family $\sigma = \{\sigma_x\}$ in which σ_x maps blind views to $\mathcal{Y}$ and is computed from the corpus with item x deleted: it may therefore use the corpus's questions, labels and other items' answers, but never a_x. For a class $\mathcal{C}$, $\beta^\star(\mathcal{C}) = \sup_{\sigma \in \mathcal{C}} N^{-1} \sum_x \mathbf{1}[\sigma_x(q_x, \ell_x) = a_x]$.

The estimand is an *auditable blind-exploitability lower bound under an explicitly enumerated strategy class* (the accuracy of the best *named* strategy in a family S printed in full), and not the fraction of a corpus "inherently" solvable without the modality. There is no unrelativised $\beta^\star$ to target: on a fixed finite corpus the supremum over all blind strategies is 1, because a family may assign σ_x separately for each deletion and choosing the constant map to a_x never reads a_x *through* $D \setminus x$ (Appendix B). A quantity that equals 1 on every corpus is not a target.

Proposition 1 (One-sidedness)**.** *Let S be a finite, explicitly enumerated set of blind strategies and $\hat{\beta} = \max_{\sigma \in S} N^{-1} \sum_x \mathbf{1}[\sigma_x(q_x, \ell_x) = a_x]$. Then (i) $\hat{\beta}$ is* attained*: some named $\sigma \in S$ is correct on that share of the corpus, computed from the annotations alone, so a large $\hat{\beta}$ condemns and the condemnation is exhibited rather than inferred; (ii) $\hat{\beta} \le \beta^\star(\mathcal{C})$ for every $\mathcal{C} \supseteq S$, so enlarging the family can only raise the bound; (iii) $\hat{\beta}$ cannot acquit: whatever S is, provided $|\mathcal{Y}|$ exceeds the number of answer-independent rules in it, there is a corpus on which $\hat{\beta} = 0$ while a blind strategy outside S is correct on every item.* ■

The asymmetry is not a defect of this instrument but of the question: an instrument that acquitted would have to quantify over every way of exploiting a corpus without looking at it, including ways nobody has thought of.

Proposition 2 (Announcement is a distinct exposure detector, not independent evidence)**.** *Let V be the share of items whose gold answer's content tokens all occur in the question. On an announced item the gold lies in the set of sub-multisets of the question's content tokens, which is finite and enumerable from the blind view alone, so V and $\hat{\beta}$ are two counts of one failure and their sum is not a count of anything. It is* not *a special case in the strategy sense: containing the gold in an enumerable candidate set is not selecting it. No inequality between the two rates follows, and none holds: a corpus can announce every option in every question, which* VSI-BENCH *(choice) does at $V = 100.00\%$ against $\beta_{g=0} = 24.94\%$.* ■

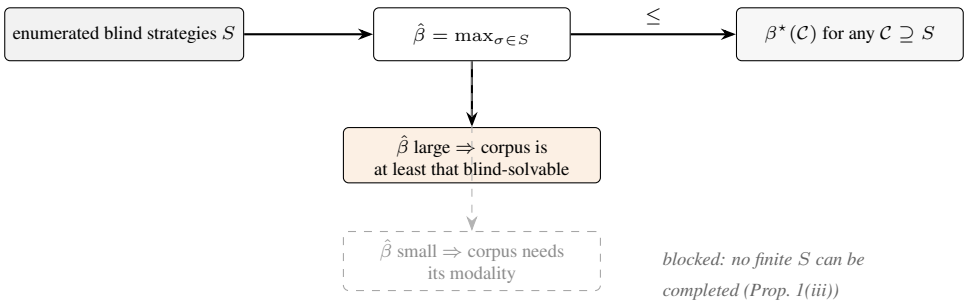

Figure 2: Proposition 1 as a diagram. The instrument's maximum over an enumerated family is a lower bound on the supremum over any wider class of blind strategies containing it, so the inference runs in one direction only. There is no unrelativised supremum to bound: over an unrestricted class it is 1 on every corpus (Remark 1). Making S larger raises $\hat{\beta}$ and therefore strengthens a condemnation, which means the incompleteness of S is in a known direction, the one property that lets a loose bound still be useful.

Definition 2 (Template, granularity, profile)**.** A template τ_g maps an item's blind view to a key, with τ_0 the corpus's own label ℓ and each increment adding one leading question token. The *profile* at g is the multiset of group sizes $\{n_i\}$ it induces.

Theorem 1 (Null-induced baseline, and collapse)**.** *Let $\hat{\beta}_g$ be the leave-one-out modal estimator at granularity g and ν_g the expectation of the same estimator when the answers are permuted uniformly at random across the corpus, holding the profile fixed. Then (i) $\nu_g > 0$ whenever the corpus's answer multiset repeats some value and some positive-probability deal gives a group a strict mode of multiplicity at least two, so $\hat{\beta}_g$ carries a null-induced positive baseline, and $\hat{\beta}_g - \nu_g$ is the* excess over the permutation null*; (ii) ν_g depends only on the profile and the corpus-wide answer multiset, not on any template–answer association, so it is computable by permuting the observed answers with no parametric model; (iii) as g increases every group size is non-increasing, once all $n_i = 1$ both $\hat{\beta}_g$ and ν_g are zero, and if some group at $g = 0$ has a strict mode of multiplicity at least two then $\hat{\beta}_0 > 0$, hence $\hat{\beta}_g$ is not non-decreasing in g: whatever else it does, it must fall to zero.* ■

Part (iii) says the curve ends on the floor. It does *not* say the curve rises first, and a sequence that starts positive and ends at zero may decrease throughout. The rise is a possibility, not a consequence, and separating the two is exactly what a reader of a single blind-baseline number needs: Proposition 7 (Appendix B) constructs an eight-item corpus that rises before it collapses, and Table 1 reports which of our rows do.

Definition 3 (Evidence leakage)**.** An estimator of blind solvability *leaks* on item x if the evidence from which it forms its prediction for x contains x's own answer. A leaking estimator is not a biased estimator of blind solvability; it estimates a different quantity, namely how well the corpus's answers predict themselves.

Proposition 3 (Two leaks and an overfit, by one mechanism)**.** *Three estimators misreport blind solvability, to degrees that grow as groups shrink; the first and third* leak *in the sense of Definition 3 and the second does not. (1)* In-sample*: each item is scored against a mode computed from a multiset containing it, so on a group of size* 1 *the item* is *the mode and the estimate tends to* 100% *as refinement isolates items. (2)* Small-group overfit*: leave-one-out at fine granularity removes the item's own answer, but in a group of size* 2 *the surviving item alone defines the mode, so the estimate measures the group's homogeneity, fitted to a sample of size one. (3)* The nonparametric bootstrap*: resampling n from n with replacement gives each item multiplicity $\sim \mathrm{Binomial}(n, 1/n) \to \mathrm{Poisson}(1)$, so conditioned on appearing at all the probability of appearing at least twice tends to $(1 - 2e^{-1})/(1 - e^{-1}) \approx 0.418$, and every extra copy sits in the held-out item's own group voting for its answer. Bootstrapping a leave-one-out estimator therefore reinstates the in-sample leak inside the uncertainty quantification.* ■

The third is dangerous precisely because it wears the costume of rigour: an author who knows to hold items out, and then attaches a confidence interval by the standard method, has undone the correction and will report the result as the more careful one.

Definition 4 (Split estimator)**.** For each group of size $n \geq 2$, partition uniformly at random into a fit half of $\lfloor n/2 \rfloor$ and a test half, compute the mode of the fit half and score the test half against it, item-weighted and averaged over partitions. Groups of size 1 admit no partition and are excluded; *coverage* is reported both by groups and by items.

Proposition 4 (The split estimator does not leak, at the price of coverage)**.** *No test item's answer enters the multiset from which its prediction is formed, so the estimator does not leak at any group size. It is unbiased for the expected item-weighted accuracy θ_g of a mode fitted on the fit half, and for nothing else, θ_g being a property of the split fraction as well as of the corpus, and* not *the accuracy of a mode fitted on $n-1$ items, which is what leave-one-out targets. Coverage is $|\{i : n_i \geq 2\}|/G(g)$ by groups, non-increasing in g, reaching 0 when every group is a singleton, and the covered subpopulation is selected by having a common template.* ■

The option regime. When answers are options rather than free text the modal-answer estimator returns almost nothing, because an option's full text is nearly unique per item, a property of the estimator, not of the corpus. For a K-way corpus we therefore take the label space to be the option index and S to contain stated surface rules. Five were fixed before any option corpus was read (INDEX, LONG, SHORT, OVERLAP, UNIQ) and four were added afterwards, each an artefact reported independently in the multiple-choice literature (FIXED, CENTROID, OUTLIER, QLENMATCH). All 9 are evaluated on every option corpus and reported in full, including those below chance; the confirmatory test uses the frozen five only, and the widening is exploratory sensitivity (Appendix B).

3 The Instrument, and What It Costs

The instrument reads question text, gold answer and grouping label, and nothing else: no frames, no model, no judge, no network. Announcement is a content-token containment test with a stoplist frozen before any corpus was read; in the option regime it is reported as V^{disc}, discriminating announcement, because a corpus that announces every option announces nothing. The blind bound is the maximum of the enumerated family, with leave-one-out scoring and a fixed blind-visible tie-break, so the maximum is attained by one named single-output strategy rather than by an oracle choice among tied modes. The permutation null is computed by re-dealing the corpus's own answers at fixed profile, $R = 1000$ draws by default. The whole audit is CPU-only and linear in R: just under two hours for 47863 items on one core, about 50 seconds for a single corpus of a few thousand items (Appendix C).

4 Results

4.1 Eight corpora, and what survives the move

Table 1 is the audit. Every option-regime corpus exceeds chance under the frozen five-rule family as a point estimate, which is what a lower bound gives. But a maximum over even five rules is a selection, and the honest test prices that selection into the null: a family-wise max-statistic permutation test over the frozen family, Holm-corrected across corpora, leaves 3 of 5 standing (Table 2). The two that fail are instructive rather than embarrassing: one owes its headline excess to its own option-count mix rather than to any blind rule, and the other's positional rule sits inside selection noise.

4.2 A positive control, on a corpus whose defect was known first

An instrument that only ever reports numbers on corpora whose defects are unknown cannot be validated, and we think this is why annotation-side audits have not displaced model-side ones. Ours can be. The parent corpus of this programme (Anonymous authors, 2026c) shipped in two versions whose only difference is a defect found by hand and documented before this instrument existed: in the earlier version every order-comparison item carried the gold BEFORE, because the generator named the earlier of two objects first. The repair assigns a balanced four–four split by a fixed permutation and leaves frames, singletons, tracked classes and chains bit-identical.

Run on both, the instrument reports $\beta = 100.00\%$ on the defective items and 0.00% after the repair, with the corpus-wide excess over the permutation null falling from 12.13 to 3.14 points. It was not

corpus	items	V / V^{disc} %	$\beta_{g=0}$	$\beta_{g^\star}$	$g^\star$	excess	option regime
OpenEQA	1636	5.75	11.67	6.72	4	+5.33	—
EgoSchema	500	0.00 / 0.00	0.80	0.80	0	+0.00	K=5: 27.40 vs 20.00 (OVERLAP)
VSI-Bench (choice)	2490	100.00 / 0.00	24.94	24.94	0	+13.39	K=4: 28.88 vs 25.00 (FIXED)
VSI-Bench (numeric)	2640	0.00	13.75	13.75	0	+2.31	—
Video-MME	2700	2.52 / 0.07	3.11	2.67	2	+2.01	K=4: 28.22 vs 25.00 (INDEX)
TempCompass	1579	0.51 / 0.06	4.50	4.50	0	+3.24	K=3: 34.39 vs 33.33 (LONG)
PerceptionTest	19140	6.80 / 4.98	21.35	45.17	8	+38.72	K=3: 36.80 vs 33.33 (QLENMATCH)
ActivityNet-QA	8000	1.19	24.62	22.56	4	+13.23	—
NExT-QA (open)	9178	1.69	6.20	6.58	2	+5.02	—

Table 1: The instrument on 8 corpora in 9 rows (VSI-BENCH enters twice, once per answer regime): two rows from the original study, above the rule, and 7 added afterwards. V is the announcement count; on the option rows the discriminating count V^{disc} of Definition 12 is printed beside it and is the one to read, because a stem that enumerates its own options announces every option at once. $\beta_{g=0}$ is the free-form blind bound at the corpus's own category granularity. $g^\star$ is the granularity at which the excess over the permutation null is largest and $\beta_{g^\star}$ is the bound there; it is *not* the maximum of β over the sweep, which on EGOSCHEMA, VSI-BENCH (CHOICE), VSI-BENCH (NUMERIC), VIDEO-MME and TEMPCOMPASS is $\beta_{g=0}$ itself and on OPENEQA is 11.86 at $g = 1$. 'excess' is over the permutation null at $g^\star$, estimated by $R = 1000$ uniform re-deals of the corpus's own answer multiset (§3). The last column is the option-regime bound against chance $1/K$ over the widened family of 9 rules, with the strategy that attained it, and is blank on the 4 free-answer rows; widening the family beyond the five rules of Definition 10 moved that bound by 0.00, 0.17 or 0.33 points depending on the corpus. This column is a *bound* and is therefore reported over the widened family, which can only raise it; the *significance* verdict of Table 2 is a different object and its confirmatory half is computed over the frozen five alone. V and β flag different items (§4.6) and the two columns are detectors of one failure family, not counts to be added: a corpus can carry either without the other, and neither column acquits anything. Every number is computed from annotations alone.

Table 2: Significance inference for the option-regime column, with the confirmatory analysis and the exploratory one separated. Both halves use the same *family-wise max-statistic permutation test*: the gold is moved to a uniformly random option slot within each item, every rule is re-scored, and the statistic is the family maximum of the raw accuracy, so selecting the best rule post hoc is priced into the null rather than ignored ($R = 2000$ replicates, seed 20260914; the statistic is the raw maximum, not a maximum of studentised t-statistics, and no per-rule standardisation is performed). The two halves differ only in which family the maximum runs over. CONFIRMATORY maximises over the 5 rules of Definition 10 as frozen before any option corpus was read; EXPLORATORY maximises over the expanded family (13 scored strategies, counting the fixed-index rule once per candidate index), four of whose rules were chosen after two option corpora had been read and which therefore cannot carry a confirmatory p-value. $p_{\max}$ is the family-wise permutation p; p_{Holm} steps the five per-corpus values down across corpora, separately within each half. Widening the family can only raise a bound (Proposition 1(ii)) but it also widens the null it is tested against, which is why the exploratory β is never smaller and its p is not simply smaller. CHANCE is $100/K$ at the corpus's median option count and the Wilson interval is on the winning frozen rule's accuracy. The permutation preserves each item's own option count, so a corpus with mixed option counts carries its K profile in the null as well as in the observation: the test asks whether the observed family max is larger than a clean corpus of the same K profile would produce. The class-level verdict of §4.1 reads the confirmatory half; the exploratory half is reported as sensitivity and changes the verdict on 0 of 5 corpora.

				CONFIRMATORY (frozen 5 rules)					EXPLORATORY (expanded family)			
Corpus	n	K	chance	β	Δ	Wilson	$p_{\max}$	p_{Holm}	β	Δ	$p_{\max}$	p_{Holm}
EgoSchema	500	5	20.00	27.40	+7.40	[23.7, 31.5]	0.0015	0.0060	27.40	+7.40	0.0015	0.0060
VSI-Bench (choice)	2,490	4	25.00	28.71	+3.71	[27.0, 30.5]	0.8536	0.8536	28.88	+3.88	0.9780	0.9780
Video-MME	2,700	4	25.00	28.22	+3.22	[26.6, 29.9]	0.1234	0.2468	28.22	+3.22	0.1239	0.2478
TempCompass	1,579	3	33.33	34.39	+1.06	[32.1, 36.8]	0.0055	0.0165	34.39	+1.06	0.0060	0.0180
PerceptionTest	19,140	3	33.33	36.47	+3.14	[35.8, 37.2]	0.0005	0.0025	36.80	+3.47	0.0005	0.0025

tuned to this defect (the defect predates it), and the two versions differ in nothing else. A second control, on a repair a third party performed and documented, is in Appendix D. We regard this as the paper's most important experiment, because it is the only one where we know the answer independently of the instrument.

Table 3: A positive control with independent ground truth. The parent paper's corpus shipped in two versions; the only difference between them is a documented defect in the order-comparison items, where the generator named the earlier object first and every gold came out BEFORE. The defect was found by hand and recorded before this instrument existed, so the instrument was not tuned to it. Frames, singletons, tracked classes and chains are bit-identical across the two columns. The instrument fires at the maximum on the defective items and reports nothing on the repaired ones. The null and excess columns are the statistic §2 recommends reporting, and they are given here because that recommendation applies to us. Note that the excess on the eight order items is 0 in *both* versions: a within-group null re-deals that group's own answer multiset onto the same group, and eight identical golds re-deal to eight identical golds, so no within-group null can see a constant gold. It is β itself, and the *corpus-wide* null of the fourth row (where the other subsets' answers enter the shuffle) that register the defect. A reader who reported only the excess would miss exactly this kind of defect, which is the one qualification we attach to the recommendation.

	before repair			after repair			
Quantity	β	null	exc.	β	null	exc.	what the corpus record says
gold share, order items		100.0%			50.0%		8 before → 4 before / 4 after
β, the 8 order items	**100.00**	100.00	+0.00	**0.00**	0.00	+0.00	instrument fires, then does not
β, the whole CMP subset	**33.33**	33.33	+0.00	**0.00**	0.00	+0.00	the defect is a third of it
β, the whole corpus (g=0)	**17.50**	5.37	+12.13	**7.50**	4.36	+3.14	diluted by the other subsets
V, the whole corpus		15.00%			15.00%		unchanged: a different failure

Table 4: Ways to score the same blind rule on the same template groups, as the template is refined. IN-SAMPLE lets each item vote for its own answer and is not an estimator of anything; it reaches 97.19%. LOO holds the item out and is what the instrument reports; it rises, then collapses. SPLIT fits the mode on half of each group and scores the other half, which is what a deployed blind system would attain. The two are not computed on the same items (LOO scores the whole corpus, including singleton groups that score 0 because their held-out remainder is empty, while SPLIT can only score groups of two or more), so LOO (cov.) repeats LOO on the covered groups alone and is the column to compare SPLIT against. On that like-for-like comparison SPLIT moves 0.84 points across $g \in \{0, 1, 2, 4\}$ against 1.91 for LOO (cov.) and 5.14 for LOO over the whole corpus. The price is coverage, reported by groups and by items because the estimators are item-weighted: by g=8 only 3.1% of groups and 6.9% of items can be scored at all, which is why that row's rise is an artefact of which groups survive and not a finding.

g	groups	singletons	in-sample	LOO	LOO (cov.)	split	cov. (grp)	cov. (item)	perm. null
0	7	0	12.04	**11.67**	11.67	11.20	100.0%	100.0%	8.29
1	92	31	17.54	**11.86**	12.09	11.05	66.3%	98.1%	7.05
2	280	168	28.48	**10.27**	11.44	11.89	40.0%	89.7%	5.49
4	1008	812	67.11	**6.72**	13.35	11.78	19.4%	50.4%	1.39
8	1571	1523	97.19	**1.65**	23.89	28.67	3.1%	6.9%	0.09

4.3 TWO LEAKS AND AN OVERFIT, MEASURED

The in-sample estimator reaches 97.19% at the finest granularity on a corpus where the held-out estimator reads 1.6% on the very same items. Leave-one-out across $g \in \{0, 1, 2, 4\}$ moves 5.14 points non-monotonically, and 10.21 points across the whole sweep. And the bootstrap inflates the bound by up to 6.96×: at one granularity the interval is $[43.40, 49.57]$ around a point estimate of 6.72. We confirm the diagnosis of Proposition 3(3) rather than assert it, by scoring the same draw twice: once as drawn, once with duplicates removed and nothing else changed. The rescored draws land within 1.38 points of the point estimate and below it in 3 of 4 rows, because a deduplicated draw is a smaller subsample with smaller groups. A paper reporting that interval would be publishing an artefact of its own error bars, as rigour. The constructive half: the split estimator is flat to within 0.84 points across the granularities where it has coverage, against 1.91 for leave-one-out on those same groups and 5.14 over the whole corpus. Most of leave-one-out's apparent collapse is singleton groups entering the average, not the estimator failing; the refinement trap is substantially, though not entirely, an artefact of how a blind baseline is scored.

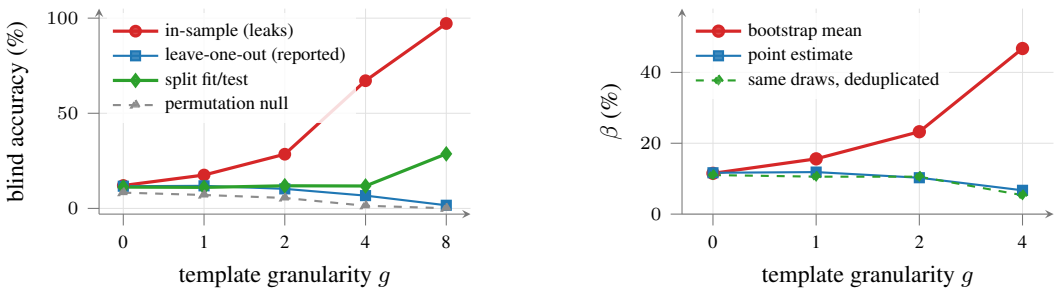

Figure 3: Left: the same blind rule on the same groups, scored four ways. In-sample rises toward 100% as refinement isolates items; the permutation null falls toward 0 for the same reason (Lemma 1); leave-one-out sits between them and is non-monotone; the split estimator is flat until its coverage collapses. **Right:** the bootstrap leak. Attaching a precision to β the standard way inflates it by 6.96× at g=4. The dashed line rescores the *same* draws with duplicates removed and nothing else changed, and it lands within 1.38 points of the point estimate (slightly below it, because a deduplicated draw is a 63% subsample with smaller groups), which identifies the cause as the duplicates rather than as the resampling.

Table 5: The instrument's mask against what a blind model actually solves. No frame, video or image was sent in any of these calls, so the models are blind by construction rather than by ablation. IN is accuracy on the items the instrument flags as blind-solvable, OUT on the rest; the comparison is within a corpus and within a model, so neither the corpus's difficulty nor the model's quality enters. Brackets are Wilson intervals. n_{in} and n_{out} sum to the items actually scored, which on OPENEQA is 1624 of 1636: the free-text rule scores containment of the gold's content tokens, and 12 golds consist entirely of stopwords (*on*, *off*) and so have none, which excludes them rather than scoring them either way, the same exclusion the announcement count makes. p_{Holm} is the two-proportion z-test on the in/out difference with Holm's step-down adjustment over the 10-test family (five backbones × two corpora); the per-test uncorrected p is in `results/a2_validate.json`. The negation rate of the gold in the IN arm is 4.4% on EGOSCHEMA and 53.1% on OPENEQA, against 2.8% and 3.7% in the OUT arm, reported so the contrast can be checked against a negation artefact rather than assumed away. This is the behavioural counterpart of the annotation-side positive control in §4.2: the control shows the instrument fires on a known defect, and this shows what it flags is what a blind system gets.

Corpus	Backbone	overall	n_{in}	IN (%)	n_{out}	OUT (%)	lift	p_{Holm}
EGOSCHEMA	Ministral 3B	25.8	137	**29.9** [22.9, 38.1]	363	24.2 [20.1, 28.9]	1.23	0.468
	Qwen3-VL-8B	33.6	137	**43.8** [35.8, 52.2]	363	29.8 [25.3, 34.6]	1.47	0.015
	Gemma 3 12B	29.2	137	**35.8** [28.2, 44.1]	363	26.7 [22.4, 31.5]	1.34	0.189
	Qwen3.8-27B	47.4	137	**46.7** [38.6, 55.0]	363	47.7 [42.6, 52.8]	0.98	0.851
	Gemma 4 31B	35.6	137	**30.7** [23.6, 38.8]	363	37.5 [32.6, 42.6]	0.82	0.468
OPENEQA	Ministral 3B	16.3	194	**51.5** [44.6, 58.5]	1430	11.5 [9.9, 13.2]	4.49	<0.001
	Qwen3-VL-8B	17.9	194	**50.5** [43.5, 57.5]	1430	13.5 [11.8, 15.4]	3.74	<0.001
	Gemma 3 12B	12.1	194	**24.2** [18.7, 30.7]	1430	10.5 [9.0, 12.2]	2.31	<0.001
	Qwen3.8-27B	18.2	194	**58.8** [51.7, 65.5]	1430	12.7 [11.0, 14.5]	4.64	<0.001
	Gemma 4 31B	11.8	194	**30.4** [24.4, 37.2]	1430	9.3 [7.9, 10.9]	3.27	<0.001

4.4 DOES THE MASK PREDICT WHAT A BLIND SYSTEM ACTUALLY SOLVES?

A positive control shows the instrument fires when it should; it does not show that what it flags is what a blind system gets, which is a claim about behaviour and requires stepping outside the instrument for the only time in this paper. We run 5 open backbones with no frame, video or image in any request, and compare accuracy inside and outside the mask, within a corpus and within a model.

On OPENEQA every backbone solves the flagged items at a higher rate, by 13.7–46.1 points, all 5 separations significant after Holm correction over the 10-test family. On EGOSCHEMA it works for weak blind models and fails for strong ones (separating on 1 of 5 backbones after Holm, and reversing on 2), and the failure is the more interesting half, because Proposition 1 predicts exactly it: the bound is a statement about an enumerated family, and a strong blind model is a strategy outside it.

4.5 AN INTERVENTION, DOSE-MATCHED

A separation is a correlation, so we intervene on it. Moving the leading token the OPENEQA mask is partly built from, with every content token left alone, does not close the gap on either backbone. We run it under two designs, the second replacing every leading token with exactly one token of the same syntactic category, so the dose is constant across items and the perturbed question stays grammatical. Both return a null, and the matched design bounds that token's contribution below roughly 20–26% of the gap rather than showing it to be zero. Decomposing the option-regime bound the other way gives the same answer from the other side: on EGOSCHEMA the larger half of the excess is positional (+5.60 against +1.80 from the score). What these corpora leak blindly is how their answers are *distributed*, not how their questions are worded, which is why paraphrasing questions is not the repair and shuffling the gold's slot at release is.

4.6 THE TWO COUNTS FLAG DIFFERENT ITEMS

On OPENEQA the announced set and the modally-predictable set intersect in 3 items where independence would predict 11.1 (a lift of 0.27, so they *avoid* each other), and together they implicate 17.42% of the corpus. Proposition 2 forbids adding them; the measurement says reporting only one would miss most of what the other finds.

5 LIMITATIONS

(i) One-sidedness is structural, and a small β acquits nothing (Proposition 1(iii) constructs the total gap). **(ii) The estimand is relative to a printed family**, and the four rules added after the first option corpus was read are marked as such throughout; the confirmatory test uses the frozen five only. **(iii) The behavioural validation is one corpus pair and 5 backbones**, and on EGOSCHEMA it fails in the direction the theory predicts. **(iv) The intervention bounds rather than excludes** the leading token's contribution. **(v) The split estimator loses coverage as g grows**, and we decline to read its value where coverage is small. Appendix E states each at length, with the recommendations that follow.

6 RELATED WORK

Shortcut and artefact detection by answer-only or hypothesis-only baselines is long established in language (Anand et al., 2018; Udandarao et al., 2025) and has been carried into multimodal evaluation by model-side audits (Majumdar et al., 2024; Lee, 2026; Xu et al., 2026); what those methods share is that they run systems. This paper computes the corpus-side part before anything runs, which is why it can be applied to corpora whose media are unavailable. The parent framework of this programme (Chen et al., 2026; Anonymous authors, 2026c) supplies the positive control and the reason the question matters: a benchmark that cannot see a capability cannot rank architectures for it, and a companion measures the corresponding arithmetic on the composition side (**?**). The estimator's failure modes connect to the standard treatment of leakage and resampling in evaluation, and the bootstrap result of Proposition 3(3) is, to our knowledge, unreported in this setting. Appendix A is the full treatment.

7 CONCLUSION

A benchmark's ability to see the capability it is cited for is partly a property of its annotations, and that part is cheap. Two counts and a curve, computed with no model and no frames, condemn a corpus by exhibiting a named strategy that solves a share of it blindly; they never acquit, and we say why no instrument of this kind can. The practical warning is that the number has no meaning without its granularity and its estimator: the same corpus reads at 1.6%, 97.2% and 11.7% under choices papers rarely state, and the ordinary bootstrap, the one an author reaches for to show care, reinstates the very leak the held-out estimator removed. We recommend the split estimator, the permutation null, and reporting both counts separately; and we recommend running the audit before release, when shuffling the gold's slot is still free.

REPRODUCIBILITY STATEMENT

Every number in this paper is written by a named analyzer to a result file, and every deciding read-out rule (the stoplist, the frozen five-rule family, the granularity sweep, the permutation budget) was fixed before the corpora it decides were read; Appendix D marks which rows were unseen at that point. Appendix B proves every statement of §2 and Appendix C gives the instrument in full. The instrument itself requires no model, no judge, no frames and no network: it reads publicly served question text, gold answers and grouping labels, so a reader can re-run the entire audit on a single CPU core without holding any corpus's media. The analyzers, the derived annotation tables and the result files are in the supplementary material.

ETHICS STATEMENT

No human subjects were involved and no annotation was collected. The instrument reads only the publicly released question text, gold answers and grouping labels of existing benchmarks, under each corpus's own licence, and never their media; the behavioural validation calls frozen open-weight models with no frame, video or image in any request. The paper's findings are limitation statements about evaluation instruments, and naming a corpus's blind-exploitable share is a criticism of a released artefact rather than of any person; we report the repair alongside each finding and, where a corpus's authors have already repaired a defect, we say so and use their repair as a control. We have read the ICLR Code of Ethics and adhere to it.

AI USE STATEMENT

We report our use of generative AI tools against the categories of the ICLR 2027 AI Policy for Authors. *Tasks in the required-disclosure list: none.* Generative AI was not used to develop the theory, to state or prove the results, to design the instrument, the null, the controls or the interventions, to set any threshold fixed before the data were read, or to interpret the results. Frozen open-weight models appear only as the *subject* of the behavioural validation of §4.4, not as tools used to produce the paper, and no model or judge is involved anywhere in the instrument itself. *Tasks in the recommended-disclosure list.* Generative AI was used on the manuscript: it condensed a longer version into the present conference format, which involved drafting and rewriting substantial parts of the exposition; it edited for readability, formatted tables and references, and assembled the anonymised supplementary archive. It was not used to discover the research topic, identify the gap, select the literature, or decide what the paper claims. *Verification.* Every quantitative statement was checked against the result file it cites. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including its text, claims and artifacts.

REFERENCES

Anonymous authors. The arbitration wall: Reliability-blindness is measurable, and it measures blind. *Manuscript under review*, 2026a. Concurrent anonymous submission; author list withheld under double-blind review. No claim in the present paper depends on it: see the paragraph "Nothing here rests on an unpublished companion".

Anonymous authors. The price of belief: What a perceptual channel yields, what it costs to check, and why checking costs what looking costs. *Manuscript under review*, 2026b. Concurrent anonymous submission; author list withheld under double-blind review. No claim in the present paper depends on it: see the paragraph "Nothing here rests on an unpublished companion".

Anonymous authors. Budget-constrained embodied perception: Four resource walls, and a pre-registered evaluation of access-structured perception on open models at 31b or below. *Manuscript under review*, 2026c. Concurrent anonymous submission; author list withheld under double-blind review. No claim in the present paper depends on it: see the paragraph "Nothing here rests on an unpublished companion".

Anonymous authors. Self-cleaning and captured anyway: One measured primitive for error in a store an agent writes to itself, and what a falling score actually measures. *arXiv preprint*, 2026a.

Anonymous authors. The structural content of a transfer relation: What a benchmark's own labels can predict about its other dimensions. *arXiv preprint*, 2026b.

Anonymous authors. Two counts over text: What a memory benchmark can see, measured before anything is run. *arXiv preprint*, 2026c.

Anonymous authors. What a rubric judge measures: Two defects in a shipped scorer, a judge that is strict and lenient at once, and the credit rate that prices the step from a count to a bound. *arXiv preprint*, 2026d.

Muhammad Qasim Ali, Saeejith Nair, Alexander Wong, et al. GraphPad: Inference-time 3D scene graph updates for embodied question answering. *arXiv preprint arXiv:2506.01174*, 2025.

Ankesh Anand, Eugene Belilovsky, Kyle Kastner, Hugo Larochelle, and Aaron Courville. Blindfold baselines for embodied QA. *arXiv preprint arXiv:1811.05013*, 2018.

Kirolos Ataallah et al. Goldfish: Vision-language understanding of arbitrarily long videos. In *ECCV*, 2024.

George I. Austin, Itsik Pe'er, and Tal Korem. Distributional bias compromises leave-one-out cross-validation. *arXiv preprint arXiv:2406.01652*, 2024.

Shuai Bai et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Anthony Brohan et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *CoRL*, 2023.

Ellis Brown, Jihan Yang, Shusheng Yang, Rob Fergus, and Saining Xie. Benchmark designers should "train on the test set" to expose exploitable non-visual shortcuts. *arXiv preprint arXiv:2511.04655*, 2025.

Boyuan Chen et al. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024a.

Liang Chen et al. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV*, 2024b.

Wenhui Chen, Jianlin Chen, Ziyao Lin, and Chi Man Vong. The capability convergence hypothesis: Capability from access structure, not scale. *arXiv preprint arXiv:2607.14144*, 2026. Public preprint, cited in the third person; the framework it states is restated in full wherever this paper relies on it.

Xueyi Chen, Keda Tao, Kele Shao, et al. Streamingtom: Streaming token compression for efficient video understanding. *arXiv preprint arXiv:2510.18269*, 2025.

Angela Dai et al. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *CVPR*, 2018.

DatologyAI, Siddharth Joshi, Haoli Yin, Rishabh Adiga, Ricardo Monti, Aldo Carranza, Alex Fang, Deng Deng, Amro Abbas, Brett Larsen, Cody Blakeney, Darren Teh, David Schwab, Fan Pan, Haakon Mongstad, Jack Urbanek, Jason Lee, Jason Telanoff, Josh Wills, Kaleigh Mentzer, Luke Merrick, Parth Doshi, Paul Burstein, Pratyush Maini, Scott Loftin, Das Das, Jiang Jiang, Vineeth Dorna, Zhengping Wang, Bogdan Gaza, Ari Morcos, and Matthew Leavitt. DatBench: Discriminative, faithful, and efficient VLM evaluations. *arXiv preprint arXiv:2601.02316*, 2026.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, et al. PaLM-E: An embodied multimodal language model. In *ICML*, 2023.

Timo Freiesleben and Sebastian Zezulka. The benchmarking epistemology: Construct validity for evaluating machine learning models. *arXiv preprint arXiv:2510.23191*, 2025.

Chaoyou Fu et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *CVPR*, 2025.

Gemini Robotics Team. Gemini robotics: Bringing AI into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Gemma Team. Gemma 4 model family (E2B/E4B/26B-MoE/31B). Technical report, Google DeepMind, apr 2026.

Kristen Grauman et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

Qiao Gu et al. ConceptGraphs: Open-vocabulary 3D scene graphs for perception and planning. In *ICRA*, 2024.

Bo He et al. MA-LMM: Memory-augmented large multimodal model for long-term video understanding. In *CVPR*, 2024.

Hsiang-Wei Huang, Wenhao Chai, Kuang-Ming Chen, et al. ToSA: Token merging with spatial awareness. *arXiv preprint arXiv:2506.20066*, 2025.

Kaixuan Jiang et al. Beyond the destination: A novel benchmark for exploration-aware embodied question answering. *arXiv preprint arXiv:2503.11117*, 2025.

Theodore J Kalaitzidis. The evaluation trap: Benchmark design as theoretical commitment. *arXiv preprint arXiv:2605.14167*, 2026.

Moo Jin Kim et al. OpenVLA: An open-source vision-language-action model. In *CoRL*, 2024.

Jae Joong Lee. Accuracy without grounding: Diagnosing visual dependency dissociation in video LLM benchmarks. *arXiv preprint arXiv:2607.13305*, 2026.

Kevin Y. Li, Sachin Goyal, Joao D. Semedo, et al. Inference optimal VLMs need fewer visual tokens and more parameters. *arXiv preprint arXiv:2411.03312*, 2024a.

Wei Li, Hehe Fan, Yongkang Wong, Mohan Kankanhalli, and Yi Yang. TOPA: Extending large language models for video understanding via text-only pre-alignment. *arXiv preprint arXiv:2405.13911*, 2024b.

Yuanxin Liu, Shicheng Li, Yi Liu, et al. TempCompass: Do video LLMs really understand videos? In *Findings of ACL*, 2024.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, et al. OpenEQA: Embodied question answering in the era of foundation models. In *CVPR*, 2024.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2023.

Microsoft. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025.

Mistral AI. Ministral 3b: A compact multimodal model. Model card, Mistral AI, 2025.

Lorenzo Pacchiardi, Marko Tesic, Lucy G. Cheke, et al. Leaving the barn door open for clever hans: Simple features predict LLM benchmark answers. *arXiv preprint arXiv:2410.11672*, 2024.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, et al. Perception test: A diagnostic benchmark for multimodal video models. In *NeurIPS Datasets and Benchmarks*, 2023.

Qwen Team. Qwen3-VL technical report. Technical report, Alibaba Group, 2025.

Qwen Team. Qwen3.8-27B: Dense, natively multimodal, Apache-2.0. Release report, Alibaba Group, aug 2026.

Santhosh Kumar Ramakrishnan et al. Habitat-Matterport 3D Dataset (HM3D): 1000 large-scale 3D environments for embodied AI. In *NeurIPS Datasets and Benchmarks*, 2021.

Krishan Rana et al. SayPlan: Grounding large language models using 3D scene graphs for scalable robot task planning. In *CoRL*, 2023.

Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. VideoRAG: Retrieval-augmented generation with extreme long-context videos. In *ACM SIGKDD*, 2026.

Manolis Savva et al. Habitat: A platform for embodied AI research. In *ICCV*, 2019.

Yuzhang Shang et al. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024.

Xiaoqian Shen, Wenxuan Zhang, Jun Chen, and Mohamed Elhoseiny. Vgent: Graph-based retrieval-reasoning-augmented generation for long video understanding. In *NeurIPS*, 2025.

Mohit Shridhar et al. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *CVPR*, 2020.

Enxin Song et al. MovieChat: From dense token to sparse memory for long video understanding. In *CVPR*, 2024.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6418–6428, 2019.

Vishaal Udandarao, Shyamgopal Karthik, Surabhi S. Nath, et al. Solving spatial supersensing without spatial supersensing. *arXiv preprint arXiv:2511.16655*, 2025.

Subin Varghese, Joshua Gao, Asad Ur Rahman, et al. BridgeEQA: Virtual embodied agents for real bridge inspections. *arXiv preprint arXiv:2511.12676*, 2025.

Yifei Wang, Zhenkai Li, Tianwen Qian, et al. StreamEQA: Towards streaming video understanding for embodied scenarios. *arXiv preprint arXiv:2512.04451*, 2025.

Rujie Wu, Xiaojian Ma, Hai Ci, et al. LongViTU: Instruction tuning for long-form video understanding. *arXiv preprint arXiv:2501.05037*, 2025.

Tao Wu et al. NoisyEQA: Benchmarking embodied question answering against noisy queries. *arXiv preprint arXiv:2412.10726*, 2024.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9777–9786, 2021.

Yige Xu, Yongjie Wang, Zizhuo Wu, et al. Do vision-language models truly perform vision reasoning? a rigorous study of the modality gap. *arXiv preprint arXiv:2604.16256*, 2026.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025a.

Jingkang Yang et al. EgoLife: Towards egocentric life assistant. In *CVPR*, 2025b.

Rui Yang et al. EmbodiedBench: Comprehensive benchmarking of multimodal LLMs for vision-driven embodied agents. In *ICML*, 2025c.

Shusheng Yang, Jihan Yang, Pinzhi Huang, et al. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025d.

Yuan Yao et al. MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9127–9134, 2019.

Haoji Zhang et al. Flash-VStream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024.

Yong Zhao et al. CityEQA: A hierarchical LLM agent on embodied question answering benchmark in city space. *arXiv preprint arXiv:2502.12532*, 2025.

Lexin Zhou, Lorenzo Pacchiardi, Fernando Martínez-Plumed, et al. General scales unlock AI evaluation with explanatory and predictive power. *arXiv preprint arXiv:2503.06378*, 2025.

A Related Work, in Full

Blind baselines, found by running them. Question-only baselines matched the state of the art on embodied question answering when the task was introduced (Anand et al., 2018). On a widely used embodied benchmark, the authors' own blind baselines score far above chance, a text-only language model reaching a substantial fraction of what the multimodal systems reach (Majumdar et al., 2024). A bag-of-words baseline discarding almost all temporal structure nearly solves a long-horizon spatial-recall task, including on multi-hour video (Udandarao et al., 2025), against a parent that proposes the task and reports collapse beyond an hour of streaming (Yang et al., 2025d). A visual-dependency diagnostic separates accuracy from grounding across twenty models and four viewing conditions, and finds frame diversity rather than temporal order supplying most of the benefit (Lee, 2026). A controlled study of the modality gap finds text-only inputs best and vision-only worst (Xu et al., 2026). The pattern is old enough to have its own name in the adjacent literature, where simple features are shown to predict benchmark answers directly (Pacchiardi et al., 2024).

All of these run systems. Ours does not, and the two kinds of evidence answer different questions: theirs is about what models do, ours about what a corpus permits. The distinction matters for what each can conclude. A model-side result showing a blind system scoring highly is evidence that the corpus is partly blind-solvable *and* that this particular model exploits it; our annotation-side result establishes the first without the second, at the cost of a CPU pass over the annotations rather than of a model sweep, and before any system has been tuned against the corpus.

Benchmark stress-testing and curation, and how they differ from auditing. A recent strand systematises the model-side diagnosis. Brown et al. (2025) propose a Test-set Stress-Test (TsT): fine-tune a strong LLM on the non-visual text inputs of the test set via k-fold cross-validation, and use the resulting score to assign each item a bias score, which an Iterative Bias Pruning procedure then filters. DatologyAI et al. (2026) take the item-curation further, converting multiple-choice to generative tasks and filtering blindly solvable and mislabeled items to produce a cleaned evaluation suite covering 33 datasets. Both ask a model-side question and a curation question: *can a model discover or exploit the shortcut, and which items should be removed?* Ours asks a third, prior question: *what is a formally auditable lower bound on blind exploitability that the corpus's own annotations certify, with a named strategy that attains it, before any model is run?* The answer is not a filtered benchmark or a bias score; it is a one-sided bound with an explicit strategy class, a fixed blind-visible tie-break, and a permutation null that separates the corpus-level signal from the exchangeability baseline. The three are complementary: TsT and DatologyAI et al. (2026) tell a practitioner what to do once a problem is found; our instrument finds the problem on annotations alone, with no model and no media, and says how much of the corpus is affected without naming a model. The difference in what each needs is also the difference in what each can be argued out of: a model-side finding depends on a particular model and its training exposure, so it can always be met with the reply that the model saw the corpus, while a bound computed from the annotations by a printed rule cannot be answered that way (§4.1). Neither subsumes the other, and the bound here is the input a filtering or stress-testing decision would want first.

The corpora, and why these two. The embodied and long-horizon question-answering literature has grown quickly and along two different answer regimes, which is why the instrument needs two estimators rather than one. Open-answer corpora, where a gold is a string, include the episodic-memory and spatial categories of OpenEQA (Majumdar et al., 2024), the spatial-reasoning items of VSI-Bench (Yang et al., 2025a), and a growing family of embodied variants that add exploration, noise, scale or a domain (Das et al., 2018; Jiang et al., 2025; Wu et al., 2024; Zhao et al., 2025; Varghese et al., 2025; Wang et al., 2025; Yang et al., 2025c; Gemini Robotics Team, 2025). Option corpora, where a gold is an index into a written option set, include EgoSchema (Mangalam et al., 2023) and the multiple-choice portions of the long-video evaluation literature (Fu et al., 2025; Wu

et al., 2025), drawing on egocentric sources (Grauman et al., 2022; Yang et al., 2025b). We instrument 8 of them across both regimes, in 9 rows, which is enough for the class-level statement of §4.1 and still not a survey of a field; §5 says which of the two this is.

What a corpus is asked to certify, and by whom. The corpora this instrument is aimed at are cited as evidence for capabilities (episodic memory, spatial reasoning, long-horizon grounding) that the systems evaluated on them are then said to have. That inferential step is what makes blind-solvability a validity question rather than a curiosity. It appears across the embodied stack: benchmarks built on scanned interiors (Ramakrishnan et al., 2021; Dai et al., 2017) and simulators over them (Savva et al., 2019; Shridhar et al., 2020), egocentric corpora assembled from worn cameras (Grauman et al., 2022; Yang et al., 2025b), and the vision-language-action systems whose progress is reported against them (Driess et al., 2023; Brohan et al., 2023; Kim et al., 2024; Chen et al., 2024a). When a corpus's headline number is partly settled before the modality is consulted, every claim that inherits it inherits the same discount, and nothing downstream can recover the amount.

Where the modality is expected to matter, and where it demonstrably does not. The architectures under evaluation increasingly do not consume raw observations at all. Streaming systems compress into a bounded bank (Song et al., 2024; He et al., 2024; Zhang et al., 2024; Chen et al., 2025); scene-graph methods maintain a symbolic abstraction (Ali et al., 2025; Rana et al., 2023; Gu et al., 2024); retrieval systems store model-written text and search it (Ren et al., 2026; Shen et al., 2025; Ataallah et al., 2024). Each of these interposes a representation between the pixels and the answer, and the parent programme measures what that interposition costs (Anonymous authors, 2026c;a;b). A blind bound is the complement of that measurement: it says how much of the task never required the pixels in the first place, which is the term every one of those cost accounts needs as a baseline and which none of them currently computes.

Backbones, and why the instrument does not name any. Every number in this paper is independent of the model that would be evaluated, which is unusual enough to be worth stating against the norm. The open multimodal backbones this programme otherwise uses (Mistral AI, 2025; Qwen Team, 2025; Gemma Team, 2025; Qwen Team, 2026; Gemma Team, 2026; Bai et al., 2025; Yao et al., 2024; Microsoft, 2025) appear nowhere in the instrument, and neither does any closed one. That is the property that lets a corpus's authors run it before a single system exists, and it is also the property that bounds what it can conclude: nothing here is evidence about any model's behaviour, only about what the corpus permits (§5, item 5).

Instruments that read a corpus instead of a model. The closest precedent in spirit is an audit of a long-term memory benchmark that counts, over the corpus text and with no model call, the share of items whose answer is announced inside a single record, together with the retrieval budget at which containment closes (Anonymous authors, 2026c). This paper carries that design into the multimodal setting, where the audited axis is the modality rather than the state-versus-index distinction, and adds the refinement correction that the text setting did not need. Simple-feature predictors of benchmark answers (Pacchiardi et al., 2024) are close in method and different in purpose: they demonstrate that surface cues suffice, without treating the estimate as a bound, without a null, and without pricing its granularity.

Benchmark validity as meta-science. Construct validity for machine-learning evaluation has an explicit literature (Freiesleben & Zezulka, 2025), including the argument that a benchmark's design is a theoretical commitment and that proxy substitution is its characteristic failure (Kalaitzidis, 2026). Predictive frameworks estimate performance from annotation-derived task demands (Zhou et al., 2025). Those accounts are conceptual or demand-based; the quantity here is a bound with a stated null and a stated one-sidedness, which is a narrower object and a checkable one. The relationship is complementary rather than competitive: a validity argument says what a corpus ought to measure, and β says how much of it is settled before the modality is consulted.

Refinement, and why more labels is not more information. That refining a label set multiplies the claims while dividing the evidence has been measured on eight text corpora, where the description length of a transfer relation splits into a base-rate term and a structural term and only the second is about transfer (Anonymous authors, 2026b). Theorem 1 is the same tension in a different estimator:

here refinement destroys a blind bound, on 5 of the 9 rows measured here after inflating it first (Proposition 7), and the permutation null is what separates the two. The general shape (an estimator that improves as its evidence thins, because it is fitting the thinning) is the classical overfitting phenomenon, and its appearance inside a *diagnostic* rather than inside a model is what makes it easy to miss.

What a verdict from an instrument is worth. Since the instrument's output is used to judge a corpus, it is subject to the same scrutiny it applies. An audit of a shipped rubric judge finds defects that a variance framing cannot express, because they change what is measured rather than how noisily (Anonymous authors, 2026d); a companion result measures how errors enter and persist in a stored representation (Anonymous authors, 2026a). Our response to that scrutiny is §4.2: a positive control on a corpus whose defect was independently documented. We are not aware of a blind-baseline method that reports one, and we think the absence is why annotation-side evidence is treated as weaker than it is.

Design-side accounts of the same phenomenon. At a fixed inference budget the optimal vision-language configuration uses very few visual tokens with the largest language model that fits (Li et al., 2024a), and the token-pruning literature shows how much of a visual input can be discarded with little loss (Chen et al., 2024b; Shang et al., 2024; Huang et al., 2025). These are statements about how to build a system, and they predict that systems will under-use a visual channel for reasons having nothing to do with corpus defects. A companion analysis in this programme adds a third account, in which a channel is down-weighted because checking it costs what using it costs (Anonymous authors, 2026b), and a fourth in which an arbiter that cannot afford to verify weights each channel by what it holds rather than by what it is worth, so a channel is down-weighted on every query it cannot address whatever its reliability elsewhere (Anonymous authors, 2026a). All four can be true at once. The instrument here is agnostic between them: it bounds what the corpus permits and says nothing about why a given system behaves as it does. What it does supply to that debate is a subtraction (the share of a corpus's headline number that was never about the modality), which every one of those accounts needs and none of them currently has.

Relation to the companion submissions, stated rather than left to be discovered. This paper belongs to a programme with an overlapping author set, and reviewing policy treats any concurrent archival submission by such a set as prior work. We therefore state the dependency structure explicitly, including what each paper is *forbidden* to claim, so that the boundaries can be checked rather than inferred. The public preprint (Chen et al., 2026) states the hypothesis that capability is bought by access structure under a fixed per-token budget. Four submissions follow, and no two share a population, an instrument and a question:

- a budgeted access-structure grid whose outcome is *accuracy*, testing whether reallocating a fixed budget beats growing it (Anonymous authors, 2026c);
- an analysis of arbitration between channels whose outcome is *which exit an access loop takes*, computed from round transcripts, which turns reliability-blindness from the assumption its own consequences rest on into a term a transcript settles, and settles it blind (Anonymous authors, 2026a);
- a model-free audit of what a benchmark's own annotations can see, which calls no model at all (this paper);
- a measurement of what it costs to *check* a perceptual claim on a metered interface, whose population is caption probes rather than agent decisions (Anonymous authors, 2026b).

This paper is the third, and the only one of the four that survives the retraction of the others: it calls no model, shares no experimental artefact with any of them, and its subject is the corpora on which their numbers are scored. It uses no number produced by a model run for any companion, because a model-free audit that borrowed one would have voided its own claim. That property is worth naming as what it is. The companions are under review and carry no archival identifier, so a claim resting on one would be a claim a reader cannot check; this paper has none. Every quantity below is computed by the released instrument from annotations the corpora themselves publish, and the companions are cited for motivation and for the questions this paper does not ask.

B THEORY, IN FULL

B.1 BLIND SOLVABILITY, AND WHAT AN INSTRUMENT CAN BOUND

Definition 5 (Corpus, and the blind view)**.** A corpus is a finite set of items $x = (q, a, \ell)$ with question q, gold answer $a \in \mathcal{Y}$, and grouping label ℓ supplied by the corpus. The *blind view* of an item is (q, ℓ): everything except the modality the benchmark is named for and except the answer. The instrument's headline estimand is an *auditable blind exploitability lower bound under an explicitly enumerated strategy class*: the accuracy of the best *named* blind strategy in a family S printed in full, not the fraction of the corpus "inherently" solvable without the modality, which no finite family can identify.

Definition 6 (Blind strategy, and blind solvability relative to a class)**.** A *blind strategy* is a family $\sigma = \{\sigma_x\}_x$ in which σ_x maps blind views to $\mathcal{Y}$ and is computed from the corpus with item x deleted: it may therefore use the corpus's questions, labels and other items' answers, but never a_x, the answer of the item it is predicting. For a class $\mathcal{C}$ of blind strategies write

$$\beta^\star(\mathcal{C}) \;=\; \sup_{\sigma \in \mathcal{C}} \; \frac{1}{N} \sum_x \mathbf{1}\big[\sigma_x(q_x, \ell_x) = a_x\big].$$

Remark 1 (Why there is no unrelativised $\beta^\star$, and why that is not a retreat)**.** It is tempting to define blind solvability as the supremum over *all* blind strategies and to treat that as the quantity the instrument bounds. On a fixed finite corpus the supremum is 1, and it is 1 for a reason that has nothing to do with the corpus. Delete item x; the remaining corpus $D \setminus x$ is, for generic x, a distinct object from $D \setminus x'$ for every other x', and a family may assign σ_x separately for each of those distinct inputs. Choosing σ_x to be the constant map to a_x never reads a_x *through* $D \setminus x$ (it reads it through the index x), so nothing in the definition forbids it, and the supremum is attained at 1 on every corpus. A quantity that is 1 for every corpus is not a target. Every claim in this paper is therefore relative to a stated class: $\hat{\beta}$ is a maximum over a family S printed in full, and one-sidedness (Proposition 1) says what that maximum licenses. This is the same reason the paper's estimators hold items out and report coverage rather than quoting a supremum: the interesting object is the accuracy of a rule someone can name, not the accuracy of the best rule that exists.

Proposition 5 (One-sidedness)**.** *Let S be a finite, explicitly enumerated set of blind strategies and let $\hat{\beta} = \max_{\sigma \in S} \frac{1}{N} \sum_x \mathbf{1}[\sigma_x(q_x, \ell_x) = a_x]$. Then*

(i) $\hat{\beta}$ *is* attained*: some named $\sigma \in S$ is correct on that share of the corpus, computed from the annotations alone. A large $\hat{\beta}$ therefore condemns, and the condemnation is exhibited rather than inferred.*

(ii) $\hat{\beta} \le \beta^\star(\mathcal{C})$ *for every class $\mathcal{C} \supseteq S$, so enlarging the family can only raise the bound, never lower it.*

(iii) $\hat{\beta}$ *cannot acquit. Whatever S is, provided $|\mathcal{Y}|$ exceeds the number of answer-independent rules in it, there is a corpus on which $\hat{\beta} = 0$ while a blind strategy outside S is correct on every item.* ■

The asymmetry is not a defect of this instrument but of the question. An instrument that acquitted would have to quantify over every way of exploiting a corpus without looking at it, including ways nobody has thought of, which is why we report only condemnations and report them as lower bounds.

Proposition 6 (Announcement is a distinct exposure detector, not independent evidence)**.** *Let V be the share of items whose gold answer's content tokens all occur in the question. On an announced item the gold lies in the set of sub-multisets of the question's content tokens, which is finite and enumerable from the blind view alone. Announcement is therefore a statement of the same kind as the blind bound (the answer is available without the modality, up to a choice inside a candidate set the question fixes), so V and $\hat{\beta}$ are two counts of one failure and their sum is not a count of anything. It is* not *a special case in the strategy sense: containing the gold in an enumerable candidate set is not selecting it, and no rule in S attempts that selection.*

No inequality between the two rates follows, and none holds in general. Choosing the correct sub-multiset requires the gold, which a blind strategy may not read; and a corpus can announce every option in every question, which is what VSI-BENCH *(choice) does at $V = 100.00\%$ against $\beta_{g=0} = 24.94\%$ in Table 1. That row is also why Definition 12 reports V^{disc} in the option regime.* ■

Proposition 2 is worth stating because it prevents a mistake the two counts invite: reading them as independent evidence. Announcement is not a separate finding from the blind bound, it is the same finding reached by a cruder route, and adding the two would double-count. What it does *not* license is an ordering of the two rates, and §4.6 reports the measured relation instead.

B.2 WHY THE NUMBER NEEDS ITS GRANULARITY

The practical estimator groups items by a template of granularity g and predicts, for each item, the modal answer of its group with that item held out. Granularity is not a free choice that washes out: it controls the group-size profile, and the profile alone fixes a positive value the estimator returns under the exchangeability null.

Definition 7 (Template, granularity, profile)**.** A template τ_g maps an item's blind view to a key, with τ_0 the corpus's own label ℓ and each increment adding one leading question token. A question with fewer than g content tokens contributes all of them and no more, so its key stops growing and its group stops splitting; on a corpus of short questions the partition therefore plateaus rather than reaching one group per item, which is visible in Table 1 as a curve that goes flat rather than to zero. The induced partition has group sizes $n_1, \ldots, n_{G(g)}$, the *profile* at granularity g.

Theorem 2 (Null-induced baseline, and collapse)**.** *Fix a corpus and let $\hat{\beta}_g$ be the leave-one-out modal estimator at granularity g. Let ν_g be the expectation of the same estimator when the answers are permuted uniformly at random across the corpus, holding the profile fixed. Then*

(i) *$\nu_g > 0$ whenever the corpus's answer multiset repeats some value and some positive-probability deal gives a group a* strict *mode of multiplicity at least two, which holds in particular whenever some group has $n_i = 2$. So, where the hypothesis holds, $\hat{\beta}_g$ is positive under the null alone and therefore carries a null-induced positive baseline when read as an estimate of the structure the template carries. If instead every answer in the corpus is distinct, $\nu_g = 0$ for every profile. We call $\hat{\beta}_g - \nu_g$ the* excess over the permutation null*; that name is a definition, and it is a modelling choice, not a further theorem, to read it as the part attributable to template–answer association, because ν_g is by (ii) exactly the value the same single-valued estimator returns when that association is destroyed and nothing else is. Lemma 1 is stated for the same single-valued rule (fixed blind-visible tie-break), so the null and the estimate are like-for-like;*

(ii) *ν_g depends only on the profile and the corpus-wide answer multiset, not on any association between template and answer, so it can be computed by permuting the observed answers, with no parametric model of either;*

(iii) *as g increases, every group size is non-increasing; once all $n_i = 1$ both $\hat{\beta}_g$ and ν_g are zero; and if some group at $g = 0$ has a strict mode of multiplicity at least two, $\hat{\beta}_0 > 0$. Hence, on such a corpus, $\hat{\beta}_g$ is not non-decreasing in g: whatever else it does, it must fall to zero.* ■

Part (i)'s hypothesis, checked on every sweep rather than assumed. The hypothesis has two conjuncts and we check both. The first holds everywhere here: all 9 of the 9 corpus rows repeat some answer. So does the second, in every cell: all 81 corpus–granularity cells behind Table 1 still have some group holding two items, no sweep having reached one group per item inside the range swept. In 76 of those 81 the measured ν_g is positive outright. The other 5 are all EGOSCHEMA at $g \geq 4$, and they are not counterexamples: ν_g there is positive but smaller than the two decimals the sweep stores. That is visible in the data rather than assumed, because the spread over draws is nonzero where the rounded mean is not (a mean of zero with positive spread is impossible), and a $20{,}000$-draw re-run puts ν_g between 0.006% and 0.001% rather than at zero. That corpus is the hard case for (i) by construction: 98.6% of its gold answers are unique, so a re-deal almost never hands a group a repeated answer. It is also why the check reports resolution separately from sign instead of collapsing the two, which an earlier version did, and which would have recorded a rounding artefact as a failure of the theorem. Where every group is already a singleton the hypothesis fails and $\nu_g = 0$; that is part (iii)'s collapse and not a counterexample to (i) either, and no cell in this sweep is yet in that regime.

Part (iii) says the curve ends on the floor. It does *not* say the curve rises first, and a sequence that starts positive and ends at zero may decrease throughout. The rise is a possibility, not a consequence, and it is worth separating the two because the difference is exactly what a reader of a single blind-baseline number needs to know.

Proposition 7 (The refinement curve can rise before it collapses)**.** *There is a corpus and an admissible template family for which $\hat{\beta}_0 = 62.5\%$, $\hat{\beta}_1 = 87.5\%$ and $\hat{\beta}_2 = 0\%$ under the single-valued modal rule with the fixed tie-break. Hence $\hat{\beta}_g$ is not monotone in g in general, and no ordering between $\hat{\beta}_0$ and $\hat{\beta}_g$ for intermediate g follows from Theorem 1.* ■

Whether a given corpus takes the rising shape or the falling one is a measurement, not a theorem, and we report it as one: of the 9 rows of Table 1, 5 have a non-monotone curve and 4 are monotone across the whole sweep: PERCEPTIONTEST rises throughout, and EGOSCHEMA, VSI-BENCH (CHOICE) and VSI-BENCH (NUMERIC) fall throughout. EGOSCHEMA falls fastest and for a reason worth naming, because it is the shape a reader should expect whenever the free-form estimator is pointed at an option corpus: its "answers" are whole option texts and are nearly unique per item, so the estimate leaves 0.80% at the coarsest granularity and is on the floor one refinement step in. The shape also depends on the template family: Appendix G cuts the same corpus by trailing rather than leading tokens and gets a curve that falls throughout, while the excess still peaks in the interior in both.

Two consequences we act on. A reported blind-baseline number is uninterpretable without the granularity at which it was computed, because both directions of error are reachable (upward by Theorem 1(i) at intermediate g, downward by (iii) at fine g), and which of them a corpus exhibits cannot be predicted from the corpus's name. And the statistic to report is the excess $\hat{\beta}_g - \nu_g$ over the permutation null together with the curve, not a single $\hat{\beta}$: the excess is the part that survives destroying the template–answer association while holding the answer multiset and group-size profile fixed, and the curve is what shows where the estimator stopped being one. It is a randomisation contrast under that null, not an identified "structural component"; the permutation destroys question–answer association along with any semantic or category-specific structure, so no unique isolation of "shortcut structure" is claimed.

This defect has a name in statistics, and the sign there is the other one. Theorem 1 is not the first statement that leave-one-out can manufacture the structure it is used to detect. Austin et al. (2024) show that leave-one-out induces a dependence between the average label of a training fold and the label of its held-out instance, term it *distributional bias*, show it persists for leave-P-out and across modelling and evaluation approaches, and propose a rebalanced cross-validation that corrects it. We claim no priority over that phenomenon and position Theorem 1 as its re-derivation in a different setting.

The setting changes the sign, and it is worth saying why, because the two directions are the same fact seen from opposite estimators. Their dependence is *negative* and it *penalises*: a model that regresses to its training mean is scored against a held-out label pulled the other way, so performance is understated and stronger regularisation looks worse than it is. Ours is *positive* and it *rewards*: a modal estimator copies its group's most common answer, and whenever the corpus's answer multiset repeats, the held-out label agrees with that mode more often than association with the template would explain; so $\nu_g > 0$ under permutation alone, and a blind bound read off $\hat{\beta}_g$ rather than off $\hat{\beta}_g - \nu_g$ counts that baseline as structure. What Theorem 1 adds to the transported phenomenon is the two things this paper needs and a correction alone would not give: a null that *prices* the artefact without a parametric model, by permuting the corpus's own answers (part (ii)), and the collapse (part (iii)) that makes the granularity sweep necessary rather than optional. Austin et al. (2024) would have us repair the estimator; we also need to report what the unrepaired one was worth, because published blind baselines are computed with it.

B.3 THE LEAK, ITS TWO ROUTES, AND THE OVERFIT BESIDE IT

Theorem 1 describes a bias without naming its mechanism. The mechanism is close to singular, and naming it explains three otherwise unrelated pathologies, one of which we did not expect to find. Two of the three are leaks in the exact sense of Definition 3 (the item's own answer is in the evidence it is scored against), and the middle one is not: it is the item's *neighbour* that decides, which is a different pathology with the same cure. We name them separately for that reason.

Definition 8 (Evidence leakage)**.** An estimator of blind solvability *leaks* on item x if the evidence from which it forms its prediction for x contains x's own answer. A leaking estimator is not a biased

estimator of blind solvability; it is an estimator of a different quantity, namely how well the corpus's answers predict themselves.

Proposition 8 (Two leaks and an overfit, by one mechanism)**.** *The following three estimators all misreport blind solvability, and to degrees that grow as groups shrink. The first and third* leak *in the sense of Definition 3; the second does not, and is named for what it does instead.*

1. In-sample *(a leak). Each item is scored against a mode computed from a multiset containing it. On a group of size* 1 *the item* is *the mode, so the estimate tends to* 100% *as refinement isolates items.*
2. Small-group overfit. *Leave-one-out at fine granularity removes the item's own answer from the evidence (so by Definition 3 it does not leak), but on a group of size* 2 *the surviving item alone defines the mode, so the estimate is the probability that two items in a group agree. That is a measure of the group's homogeneity, not the accuracy of any strategy applicable to an unseen item, and it is fitted to a sample of size one.*
3. The nonparametric bootstrap. *Resampling* n *items from* n *with replacement gives each item a multiplicity* $\sim \mathrm{Binomial}(n, 1/n) \to \mathrm{Poisson}(1)$. *Conditioned on an item appearing at all, the probability that it appears at least twice tends to*

$$\frac{1 - 2e^{-1}}{1 - e^{-1}} \approx 0.418, \tag{1}$$

and every additional copy sits in the held-out item's own group voting for its answer. Bootstrapping a leave-one-out estimator therefore reinstates the in-sample leak inside the uncertainty quantification. ■

The third is the dangerous one precisely because it wears the costume of rigour. An author who knows to hold items out, and who then attaches a confidence interval by the standard method, has undone the correction and will report the result as more careful than the uncorrected one. We measure it in §4.3 and diagnose it by scoring each draw twice: once as drawn, once with duplicates removed and nothing else altered.

Lemma 1 (I.i.d. approximation to the permutation null: where it sits, and why refinement moves it down)**.** *Let* p *be the corpus-wide answer distribution and consider the permutation null on a group of size* n *under the single-valued modal rule (ties broken by a fixed blind-visible ordering). Replace the finite-population re-deal, which samples the group's answers from the corpus's answer multiset* without *replacement, by its i.i.d. approximation: let* $X_i \overset{iid}{\sim} p$ *and write* $\nu(n) = \sum_y p_y \Pr\big[\hat{y}(X_1, \ldots, X_{n-1}) = y\big]$*, where* $\hat{y}$ *is that deterministic mode. Then*

$$\nu(1) = 0, \qquad \nu(2) = \sum_y p_y^2, \qquad \nu(n) \xrightarrow[n\to\infty]{} \max_y p_y, \tag{2}$$

and since $\sum_y p_y^2 \le \max_y p_y$*, the null is non-decreasing in group size over these three points. Consequently refining a template pushes* ν_g down *while the in-sample estimate is pushed* up*, and the two move apart for reasons that have nothing to do with the corpus's structure.* ■

The i.i.d. substitution is stated in the lemma rather than buried in its proof because it is the one place in the paper's theory where an approximation is made, and its status should be unambiguous. The null the instrument *computes* is the exact finite-population one (the corpus's own answer multiset is shuffled and re-dealt without replacement, which is what §3 means by calling it exact), and nothing downstream depends on the closed form. The lemma's role is to say *where* that simulated null sits and *why* it moves the way it does, and for that an i.i.d. expression suffices: it is exact in the large-corpus limit, and its two endpoints bracket the measurement below. Read as a finite-N identity for the without-replacement null it would be wrong; read as its asymptotic location, which is how it is used, it is what makes the collapse legible.

Lemma 1 is the closest thing this instrument has to a unit test, because both endpoints are computable from p alone and can be checked against the measurement. On OpenEQA, $\sum p_y^2 = 1.54\%$ and $\max_y p_y = 8.80\%$ over 1001 distinct answers; the measured null runs from 0.09% at the finest granularity, where almost every group is a singleton and $\nu(1) = 0$ dominates, to 8.29% at the coarsest, where groups are large and the base rate dominates. Both ends land where Eq. (2) says they must.

B.4 AN ESTIMATOR THAT DOES NOT LEAK

All three failures of Proposition 3 share a repair.

Definition 9 (Split estimator)**.** For each template group of size $n \geq 2$, partition the group uniformly at random into a fit half of $\lfloor n/2 \rfloor$ items and a test half of the remaining $\lceil n/2 \rceil$, compute the mode of the fit half, and score the test half against it. The estimate is the total number of hits over the total number of scored test items (so it is *item*-weighted, not group-weighted), averaged over random partitions. Groups of size 1 admit no partition and are excluded; the *coverage* is reported both as the share of groups and as the share of items that can be scored.

Proposition 9 (The split estimator does not leak, at the price of coverage)**.** *Under Definition 4 no test item's answer enters the multiset from which its prediction is formed, so the estimator does not leak in the sense of Definition 3, at any group size. Write θ_g for the expected item-weighted accuracy, over uniformly random half-partitions of every group of size $n_i \geq 2$, of the modal rule fitted on the fit half and applied to the test half. The split estimator is unbiased for θ_g, and for nothing else: θ_g is a property of the split fraction as well as of the corpus, and it is* not *the accuracy of a mode fitted on $n-1$ items, which is what leave-one-out targets. Its coverage is $|\{i : n_i \geq 2\}|/G(g)$ by groups and the corresponding item share by items, both non-increasing in g and reaching 0 when every group is a singleton; the covered subpopulation is selected by having a common template and is not a uniform sample of the corpus.* ■

Proposition 4 is why we report coverage in every column of the split estimator and why we decline to read its value where coverage is small. It is also the paper's one piece of constructive advice: measured across $g \in \{0, 1, 2, 4\}$, the split estimator varies by 0.84 points against 5.14 for leave-one-out over the whole corpus and 1.91 for leave-one-out restricted to the same covered groups (§4.3). The second of those comparisons is the one that is like for like, since leave-one-out over the whole corpus includes singleton groups that score zero by construction and the split excludes them; it is the smaller difference, and it is still a difference. The refinement trap is therefore substantially, though not entirely, an artefact of *how* a blind baseline is scored rather than an unavoidable property of the quantity it scores, and anyone computing a blind baseline should compute this one.

B.5 THE OPTION REGIME

When answers are options rather than free text, the modal-answer estimator returns almost nothing, because an option's full text is nearly unique per item. That is a property of the estimator, not of the corpus.

Definition 10 (Option-regime bound)**.** For a K-way corpus the label space is the option index, and S contains stated surface rules. Five were fixed before any option corpus was read: INDEX, the modal gold index within a template group, estimated leave-one-out; LONG, the longest option; SHORT, the shortest; OVERLAP, the option with most content-token overlap with the question; and UNIQ, the option with most tokens absent from the other options. Four were added afterwards, once more than one option corpus was in the table, each an artefact reported independently in the multiple-choice literature: FIXED, the best single constant index; CENTROID, the option most similar to the others; OUTLIER, its complement; and QLENMATCH, the option whose length is closest to the median. All 9 are evaluated on every option corpus and reported in full, including those below chance, and §4.1 states which corpora the widening moved and by how much. The bound is $\max_{\sigma \in S}$, which by Proposition 1 remains one-sided, and chance is $1/K$.

Each member of S is the *pair* $(\text{scoring rule}, \text{tie-break})$ and not the scoring rule alone. A surface score assigns a number to each option; where the top score is not unique (which on this family is common rather than exceptional), the score has not chosen and a fixed blind-visible ordering does. The pair is what constitutes a single-output blind strategy in the sense of Definition 1, so the pair is what the name OVERLAP or LONG denotes throughout this paper. This matters twice. Substantively, a rule's accuracy can be carried mostly by its ordering rather than by the regularity its name advertises, and on EGOSCHEMA it is (§D.11, Appendix G). Formally, it is why the bound stays one-sided: an oracle choice among tied options would not be a strategy a blind system could execute, and no member of S makes one.

Enumerating the rules in advance is what preserves the guarantee, and enlarging the family afterwards does not break it: by Proposition 1(ii) adding a rule can only raise a lower bound. What would break

it is searching a family until something exceeds chance and then reporting only the winner, so the family is fixed, small, and reported in full including the rules that fail, with the point at which each rule entered stated rather than elided.

B.6 What the number licenses, and what it does not

An instrument is only useful if its output can be put to work, and the obvious use of β is wrong. Since it is wrong in a way that looks like arithmetic, it is worth stating formally.

Proposition 10 (The subtraction is not a subtraction)**.** *Let $\sigma \in S$ attain $\hat{\beta}$ and let $A = \{x : \sigma(q_x, \ell_x) = a_x\}$ be the item set it settles, $|A|/N = \hat{\beta}$. Let a system attain accuracy* acc *on the corpus. Then* $\mathrm{acc} - \hat{\beta}$ *is not in general the share of the corpus the system solves using the modality, and it errs in both directions:*

1. *it* understates *when the system is wrong on part of A, since those items are subtracted from its credit without ever having been in it; the extreme case is a system correct only on A^c, for which the true modality-attributable share is* acc *and the subtraction reports* $\mathrm{acc} - \hat{\beta}$*;*
2. *it* overstates *when the system is correct on part of A for blind reasons, since those items remain in its credit. In the extreme case of a system that implements σ and nothing else, the true modality-attributable share is* 0 *and* $\mathrm{acc} = \hat{\beta}$ *identically, so the subtraction does report* 0*: the right answer, but only because the two rates coincide by construction, and the same subtraction reports a positive number for any system that is correct on part of A and on nothing else.*

The two agree exactly when the system is correct on all of A and correct on a further $\mathrm{acc} - \hat{\beta}$ *share of the corpus outside A, a condition on the system that no corpus-side instrument can verify.* ■

Definition 11 (What the instrument should output)**.** The instrument returns, besides the scalars, the *masks* A and $\{x : x \text{ announced}\}$, the item sets, not their sizes. The quantity we recommend reporting alongside a headline number is the conditional accuracy $\mathrm{acc}|_{A^c}$ on the items no enumerated blind strategy settles, together with $|A^c|$.

Definition 11 is why every experiment below reports item counts and why the released artefact is a mask rather than a table of rates. A conditional accuracy on a stated subset is a measurement a reader can reproduce and compare; a difference of two scalars computed under two different protocols is not. It also makes the instrument's one-sidedness operational rather than merely epistemic: enlarging S grows A, which shrinks A^c and makes the conditional accuracy a report on a harder subset, monotonically, and in the direction that cannot manufacture a favourable result.

Proposition 11 (Cost)**.** *Computing V takes $O(\sum_x (|q_x| + |a_x|))$ time. At granularity g, partitioning is $O(N)$ and the leave-one-out modal pass is $O(\sum_i n_i) = O(N)$ with a counter per group; the split estimator adds $O(RN)$ for R partitions and the permutation null $O(RN)$ for R shuffles. The whole instrument over $|G|$ granularities is therefore $O(|G|\, R\, N)$ with no dependence on any model, and it reads no media.* ■

The point of Proposition 11 is not the exponent. It is that the instrument's cost is independent of the thing that makes the model-side diagnostic expensive (the number of systems evaluated and the size of the media), so it can be run by a corpus's authors before release, on a laptop, as a matter of routine. The closest model-side diagnostic evaluates twenty models across four viewing conditions (Lee, 2026); this one loads no weights, sends no request and touches no pixel, and §C.8 states the constant that goes with the exponent. The two answer different questions and only one of them can be run before the corpus exists in the world.

C Method, in Full

C.1 What the instrument reads, and what it never touches

The instrument takes a corpus's annotation file and nothing else (Fig. 4). For each item it uses the question string, the gold answer string, and whatever grouping label the corpus supplies; for an option corpus it also uses the option strings and the gold index. It does not load frames, call a model, invoke

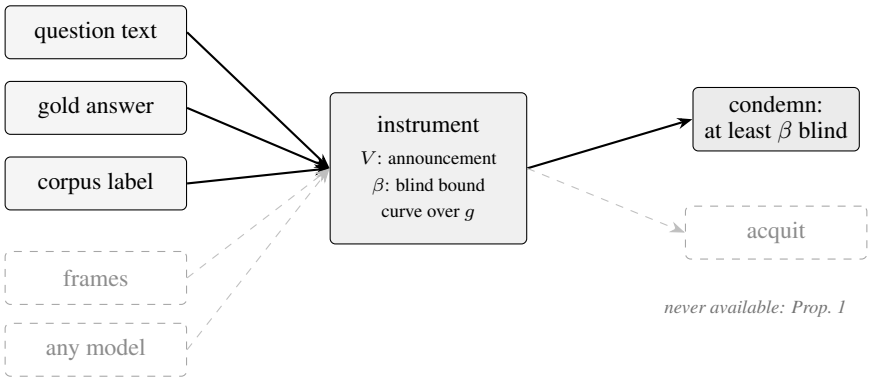

Figure 4: What the instrument reads, what it never touches, and what it can conclude. Solid arrows are inputs and available conclusions; dashed ones are deliberately absent. Frames are not read because every corpus audited here keeps them behind a signed agreement while publishing its annotations, and no model is called because the whole point is to run before anything is trained or served. Acquittal is unavailable by Proposition 1, not by omission.

a judge, or consult a retriever. That is a design requirement rather than a convenience: the pixels of every corpus audited here are behind a signed agreement, and the annotations are public, so an instrument that needed frames could not be run on the corpora that most need auditing.

Normalisation is stated in full because it is the only place a result could be manufactured. Strings are lowercased and stripped to alphanumerics and spaces. Content tokens are tokens outside a fixed stopword list, given in Appendix G. An answer whose content tokens are empty after stripping is excluded from the announcement count rather than counted either way, because matching a stopword says nothing.

Why the instrument is deliberately crude. Every component is a string operation over annotations, and each could be replaced by something stronger: a parser instead of leading tokens, a sentence embedding instead of token overlap, an entailment model instead of verbatim containment. We use none of them, for a reason that is about what the output is for rather than about effort. The instrument's guarantee is Proposition 1, drawn as Fig. 2, and that guarantee survives only while every strategy in S is one a reader can evaluate by hand on a single item. A blind bound computed with a learned component inherits that component's failure modes; and the failure modes of learned components used as measuring instruments are precisely what the surrounding literature keeps finding (Anonymous authors, 2026d; Pacchiardi et al., 2024; Kalaitzidis, 2026). It would also reintroduce the cost the instrument exists to avoid: a model in the loop is a model to serve, to version and to license, and the corpora that most need auditing are the ones whose media cannot be shipped (Ramakrishnan et al., 2021; Dai et al., 2017; Grauman et al., 2022; Yang et al., 2025b).

There is a second reason, narrower and more practical. Every quantity here is meant to be recomputable by a corpus's own authors, at release, without a GPU. A crude instrument that a maintainer will actually run is worth more than a careful one they will not, and the audit literature this paper sits in has consistently found defects that no sophisticated method was needed to find (Anand et al., 2018; Udandarao et al., 2025; Lee, 2026; Anonymous authors, 2026c).

C.2 The announcement count

V counts items whose gold answer's content tokens all appear among the question's tokens. The conjunction matters: a question that happens to contain one word of a three-word answer is not announcing it. By Proposition 2 announcement is a distinct exposure detector of the same *kind* of blind failure as β (the answer is available without the modality), so the two counts must not be added. It is *not*, however, a component of β: containing the gold in an enumerable candidate set is not selecting it, and no rule in S attempts that selection, so $V \leq \beta$ is not implemented and is not claimed.

Definition 12 (Announcement, and discriminating announcement)**.** For an item with question q and gold a, write $\mathbf{1}[a \sqsubseteq q]$ for the event that every content token of a occurs among the tokens of q. Then

$$V \;=\; \frac{1}{n}\sum_{i} \mathbf{1}[a_i \sqsubseteq q_i], \qquad V^{\mathrm{disc}} \;=\; \frac{1}{n}\sum_{i} \mathbf{1}[a_i \sqsubseteq q_i] \prod_{o \,\in\, \mathcal{O}_i \setminus \{a_i\}} \big(1 - \mathbf{1}[o \sqsubseteq q_i]\big),$$

where $\mathcal{O}_i$ is the item's option set where it has one. V is the count reported for a free-answer corpus. V^{disc} is the count reported in the option regime, and the product is the whole content of the amendment: it requires that the gold be announced *and no distractor be*.

The second clause is not a refinement, it is a correction, and §4.1 is where it became necessary. In the option regime the label space *is* the option set, so every gold appears in it by construction; a corpus whose stem enumerates its own options therefore scores $V = 100\%$ on every item, gold and distractor alike, while telling a blind reader nothing. V and V^{disc} coincide whenever the stem announces at most one option, which is every free-answer corpus and most option-regime ones, so nothing measured before the amendment changes.

C.3 The blind bound, and its null

At granularity g the corpus is partitioned by τ_g. Within each group the estimator predicts, for each item, the single-valued modal answer of the group with that item removed: the most frequent remainder answer, with ties broken by a fixed blind-visible rule (lexicographically smallest normalised answer string; smallest gold index in the option regime). The prediction is therefore exactly the output of one named single-output blind strategy, and a hit is that strategy being correct, not "gold in the mode set", which no single strategy can attain when different tied items need different favourable breaks. Holding the item out is what makes the number an estimate rather than a description: an in-sample modal predictor scores near one on any corpus with distinct answers and measures nothing.

The null is a permutation of the corpus's own answers, not a model. The corpus-wide multiset of answers is shuffled and re-dealt to the same group-size profile, and the same single-valued estimator is run. Permuting rather than assuming a parametric answer distribution is what we mean when we call the null *exact*: it is exact for the observed answer multiset, in that nothing about the answer distribution is fitted or assumed. The *expectation* over re-deals is not computed in closed form. It is estimated by R uniform shuffles and averaged, which gives ν_g of Theorem 1, and the reported statistic is the excess $\hat{\beta}_g - \nu_g$ over the permutation null. The Monte Carlo spread of that average is small and quantified by a convergence run on OpenEQA (`results/a2_nullconv.json`): at g=0 the null mean moves from 7.73 at $R = 20$ (SE 0.25) to 8.29 (SE 0.03) at $R = 1000$, a shift of 0.56 points (2.2× the $R = 20$ SE) that is concentrated at the coarse granularities; at g=8 it is 0.14 (SE 0.03) vs 0.09 (SE < 0.01). The word for that shift is *Monte Carlo error*, not bias: the average of R uniform re-deals is an unbiased estimator of the permutation expectation at every R, and what a single $R = 20$ run buys is a realised deviation of the size its own standard error predicts, not a systematically displaced estimand. One convergence run cannot establish a bias and we do not claim one. What it does establish is that at $R = 20$ the realised error is large enough to matter, because the shift is of the same order as the smallest corpus-level excesses reported here; so the main results use $R = 1000$ throughout (the budget at which the Monte Carlo SE is below 0.05 points at every granularity) rather than the $R = 20$ budget whose instability the convergence run quantifies. The other two Monte Carlo budgets in the instrument are 40 random partitions for the split estimator (§C.5) and 200 draws for the bootstrap diagnosis of §4.3.

One thing this excess is deliberately not given, and the omission is a decision rather than an oversight. $\hat{\beta}_g - \nu_g$ is reported as an *audit statistic* (a contrast between what the estimator returns on the corpus and what it returns on the corpus's own answers re-dealt), and no significance label is attached to it on any free-answer row. Doing so would require a randomisation p-value per corpus and per granularity and a correction over the 9 rows and nine granularities that produced them, and the quantity the paper uses it for does not need one: it is read as a magnitude beside its null, in a column that also prints the null. Where a *decision* is made from a comparison (which option corpora exceed chance, which backbones separate the mask's two arms), the test and its correction are supplied and stated (Tables 2 and 5). We would rather have a statistic with a stated null and no verdict than a verdict whose family we never wrote down.

C.4 THE OPTION REGIME, AS IMPLEMENTED

For a K-way corpus all 9 rules of Definition 10 are evaluated exactly over the whole corpus and the bound is their maximum; that widened family is what the option column of Table 1 reports, and §4.1 states per corpus what the widening changed. The family is reported in full, including rules that fall below chance, because a bound obtained by searching until something exceeded chance and then printing only the winner would be a selection effect. The modal-index rule is estimated leave-one-out inside template groups like the open-answer estimator; the surface rules are deterministic functions of an item and need no held-out treatment, and FIXED is a single constant index chosen once for the whole corpus.

Each surface rule scores every option and takes the argmax, and the tie-break is part of the rule rather than an implementation detail, so we state it. A rule whose top score is tied resolves toward the *largest* option index; QLENMATCH and the modal INDEX rule resolve toward the smallest. Both orderings are fixed and blind-visible (an option's position is part of the blind view), so each rule is a single-output blind strategy either way and Proposition 1(i) holds under either. Which ordering each rule uses was frozen with the rule and is not a choice made after seeing a corpus. It is not, however, cost-free: on this family a top tie always spans at least two indices, so wherever ties are common the ordering is doing part of the work the rule's name claims for its score. Appendix G measures how much, per rule and per corpus, and §5 says what that does and does not change.

C.5 THE SPLIT ESTIMATOR, AND REPORTING COVERAGE

Definition 4 is implemented as stated: within each group of size $n \geq 2$ the answers are shuffled, the first $\lfloor n/2 \rfloor$ form the fit half, the mode of that half is taken (ties broken by first occurrence, which is arbitrary but fixed), and the remaining $\lceil n/2 \rceil$ answers are scored against it. The procedure is repeated over 40 random partitions and averaged, and the standard deviation across repetitions is reported so that the Monte Carlo error is separable from the effect. Hits and scored items are pooled across groups before dividing, so the estimate is item-weighted, matching Definition 4 and the leave-one-out estimator it is compared against.

Coverage is reported in the same column. It is not a diagnostic detail: by Proposition 4 the split estimator is unbiased *on the groups it can score*, and those groups are exactly the ones with at least two items, which at fine granularity is a small and non-uniform subpopulation. Where coverage falls the estimate becomes a statement about a shrinking, self-selected part of the corpus, and §4.3 declines to read it there.

C.6 THE POSITIVE CONTROL

An instrument that reports numbers only on corpora whose defects are unknown cannot be validated, and we regard the absence of positive controls as the main reason annotation-side audits are treated as weaker evidence than model-side ones. We supply one.

The parent corpus of this programme (Anonymous authors, 2026c) shipped in two versions. In the earlier one, the generator constructed order-comparison items by naming the earlier of two objects first, so every gold came out BEFORE; a blind agent scores perfectly on that subset, and the defect was found by hand and written into the parent paper's deviation log before the present instrument existed. The repair derives a fixed permutation from the corpus seed and assigns four BEFORE and four AFTER, changing nothing else: the frames, the singleton objects, the tracked classes and the chain items are bit-identical across the two versions, which we verified rather than assumed.

The instrument is run on both versions under exactly the settings used for every other corpus in this paper: the corpus's own subset label as the coarsest template, no tuning, no special casing. Two properties make this a control rather than a demonstration. The ground truth is independent of the instrument, since the defect is documented and its mechanism is known from the generator's source. And the direction is pre-specified: the instrument must fire on the earlier version and not on the later one, and it has no free parameter with which to arrange that.

Algorithm 1 The instrument. Annotations only: no frame is opened, no model is called, and no network request is issued. The two repetition budgets are separate and are named separately: $R = 1000$ for the permutation null and $R_{\text{spl}} = 40$ for the split estimator's random partitions; the bootstrap of §4.3 is a diagnosis of what goes wrong, not a step of the instrument, and does not appear here.

Require: items $\{(q_i, a_i, \ell_i)\}_{i=1}^{N}$, granularities G, permutation repetitions R, split repetitions R_{spl}

$V \leftarrow \frac{1}{N}\big|\{i : \text{content}(a_i) \subseteq \text{tokens}(q_i),\ \text{content}(a_i) \neq \varnothing\}\big|$ ▷ announcement
for all $g \in G$ **do**
  partition items by τ_g, giving groups $\{P_1, \ldots, P_{G(g)}\}$
  $\hat{\beta}_g \leftarrow$ leave-one-out modal accuracy over all groups ▷ no item votes for itself
  $\text{ins}_g \leftarrow$ the same rule *without* holding the item out ▷ reported to expose the leak, never as an estimate
  $\text{spl}_g \leftarrow$ mean over R_{spl} random fit/test halves; record coverage ▷ Def. 4
  for $r = 1$ to R **do**
    reshuffle the corpus-wide answer multiset onto the *same* group-size profile
    $\nu_g^{(r)} \leftarrow$ leave-one-out modal accuracy on the reshuffled assignment
  end for
  $\nu_g \leftarrow \text{mean}_r\, \nu_g^{(r)}; \quad \text{excess}_g \leftarrow \hat{\beta}_g - \nu_g$
end for
if the corpus is option-regime **then**
  evaluate every rule in the fixed family S of Def. 10 exactly; $\hat{\beta} \leftarrow \max_{\sigma \in S}$
  report *all* rules, including those below chance $1/K$ ▷ a searched family would not be a bound
end if
return V; the curves $\hat{\beta}_g, \nu_g, \text{excess}_g, \text{ins}_g, \text{spl}_g$; and the profile $\{n_i\}$ at each g

C.7 What is deliberately not done

Three omissions are choices and are listed so they are not read as oversights.

No confidence interval on β. By Proposition 3 the ordinary nonparametric bootstrap reinstates the leak the estimator exists to avoid, and §4.3 measures how much. The permutation null already reported plays the role a confidence interval would: it is exact for the observed answer multiset, and it resamples no items at all. Where a precision is genuinely needed, the Monte Carlo spread of the split estimator is reported instead. The prohibition is on intervals for β, an accuracy of a rule fitted inside the same groups it is scored on, and not on intervals in general: the accuracy of a *model* on a stated item set is an ordinary binomial rate over items chosen without reference to the model, and Table 5 attaches Wilson intervals to those without inconsistency.

No search over templates. The template family is fixed in advance (the corpus's own label, then leading question tokens, with trailing tokens as a stated alternative), and both families are reported at every granularity. Searching templates until $\hat{\beta}$ exceeded a threshold would convert a bound into a selection effect.

No acquittals. By Proposition 1 the instrument cannot certify that a corpus requires its modality, and we report no such conclusion for any corpus, including the ones whose numbers are small.

C.8 Cost, measured

The number in Proposition 11 is an exponent, so it is worth stating the constant as well, and stating it once rather than describing the cost loosely in several places.

The property that distinguishes this instrument from a model-side diagnostic is categorical rather than numerical: it loads no weights, sends no request, reads no frame and needs no GPU. Within that, the cost is a CPU cost like any other, linear in the number of items N, in the number of granularities $|G|$ and in the permutation budget R. The full sweep behind Table 1 (all 9 rows, 47863 items, nine granularities each, with $R = 1000$ re-deals at every granularity) took just under two hours of single-core CPU on the ordinary workstation used here. A single mid-sized corpus is a much smaller number: the same nine-granularity sweep at the same budget takes about 50 seconds on OpenEQA's

1,636 items and about 40 seconds on EGOSCHEMA's 500. Linearity in R is what connects the two figures to the earlier budget: at $R = 20$ the whole audit ran in under a minute, and the two-hour figure is the price of retiring the Monte Carlo error that budget left (§3). We state the two-hour number rather than a favourable one because an auditing instrument's cost claim should be the one its own full protocol incurs.

The bootstrap diagnosis of §4.3 (200 draws, each scored twice) is a diagnostic run once rather than part of the instrument, and the behavioural validation of §4.4 is the one part of this paper that costs interface time rather than CPU. For contrast, the closest model-based diagnostic evaluates twenty models across four viewing conditions (Lee, 2026). The two are complements: that measurement answers a question about models, this one about corpora, and only the second can be run before a corpus ships.

D EXPERIMENTS AND ANALYSIS, IN FULL

D.1 CORPORA

8 corpora in 9 rows (VSI-BENCH enters twice, once per answer regime), 47863 items, chosen for what they contest rather than for size. Two carry the main development of the paper; 6 corpora, contributing 7 rows, were added afterwards, because a class-level claim about benchmarks cannot rest on the two corpora an instrument was designed against. Table 1 is the whole of it, and Algorithm 1 is what was run on each.

OPENEQA (Majumdar et al., 2024) supplies 1,636 items with free-text answers and 7 category labels; its frames index HM3D and SCANNET and are not used. EGOSCHEMA (Mangalam et al., 2023) is 5-way multiple choice over long egocentric video; its public answer subset gives 500 labelled items and its frames index a gated source.

The 6 added are VSI-BENCH (Yang et al., 2025a), 2490 choice items and 2640 numeric ones over indoor video, and one of the three columns the parent paper of this programme reports against, so an instrument that condemns it condemns our own evaluation; VIDEO-MME (Fu et al., 2025), 2700 items, a widely used general video benchmark in this literature; TEMPCOMPASS (Liu et al., 2024), 1579 items, built specifically to require temporal structure, which makes a positive reading there sharper than one on a corpus with no such intent; PERCEPTIONTEST (Pătrăucean et al., 2023), 19140 items whose option sets include an explicit *I don't know* and were written to defeat the surface rules the option-regime bound enumerates; ACTIVITYNET-QA (Yu et al., 2019), 8000 open-answer items over untrimmed web video, added because the free-answer estimator rested on fewer corpora than the option-regime bound did; and the open-ended split of NEXT-QA (Xiao et al., 2021), 9178 items whose questions are written around causal and temporal relations rather than around objects. The last two are open in *format* and substantially binary in *content*, which the blind bound registers as a high base rate; that is a property of those corpora and is reported rather than corrected for (Appendix G).

The regimes matter and are never mixed. The open-answer estimator and the option-regime bound are different instruments, and VSI-BENCH ships both regimes in one file (some items carry an option list and the rest a bare number), so it enters as two corpora. Averaging them would report a single β against a chance level that is 25% for half the items and undefined for the other half.

Every annotation file is public and none of the media is touched. `analysis/a2_fetch.py` retrieves them and validates each download against the serving endpoint's own row count. That check is not defensive programming but a response to two intake failures the audit trail records, each of which produced an empty or truncated corpus that parsed cleanly and raised nothing; Appendix G gives both in full, because an auditing instrument whose own intake can fail silently is not one.

D.2 EIGHT CORPORA, AND WHAT SURVIVES THE MOVE

Table 1 is the instrument on all of them. 7 of the 9 rows are corpora it had never seen when the free-form estimator, the permutation null and the refinement analysis were fixed, which is the only sense in which a measuring instrument can be said to have been tested. Two components were *amended* rather than tested by those rows, and we mark them as such wherever they appear: the announcement count acquired its discriminating form (Definition 12) after VSI-BENCH, and the

option-regime family grew from five rules to 9 (Definition 10). Neither amendment can lower a bound (Proposition 1(ii)), but neither is an out-of-sample test of the thing it amended.

An LLM with no video already scored higher than our bound, and that is the point. On EGOSCHEMA the instrument returns 27.4% against a chance of $1/5$. A reader who knows the benchmark will immediately object that published *blind* runs are higher: Li et al. (2024b) report, for models given only the question and its options, GPT-4-Turbo at 30.8%, Bard at 33.2% and Gemini-Pro-1.0 at 38.2%. We should confront those numbers rather than leave them out, and there is no contradiction to resolve: β is a *lower* bound (Proposition 1), so a language model finding eleven points more than a printable rule finds is exactly what a lower bound permits. But the honest question is sharper than consistency: what did the instrument buy, if a blind LLM already reported higher?

Three things it cannot buy, and one it can. It does not buy the largest number; a model with a world prior will always be able to exceed a rule we can write down. It does not buy an estimate of blind solvability; neither does the LLM run, which is also a lower bound. And it does not buy an explanation of the LLM's eleven points, which may be corpus structure or may be pretraining exposure to the underlying videos, and the two are not separable by that experiment. What it buys is that 27.4% is attained by a rule that reads *annotations only* (no model, no API call, no frame, no weights whose training set overlaps the corpus), so it is a property of the corpus rather than a joint property of the corpus and a particular model, it can be run before the corpus is released and on one CPU core in well under a minute at this corpus's size, and it cannot be argued away as contamination. A blind-LLM number and a blind-rule number answer different questions; the second is the one a benchmark designer can act on, and the first is the reason they should want to.

The class-level result, and the two statements it is made of. Two statements have to be kept apart here, because they are not the same statement and the first is the weaker one.

As a *point estimate*, every one of the 5 option-regime corpora sits above chance already under the five-rule frozen family: the excess over $1/K$ runs from 1.06 points to 7.40 and is positive on 5 of 5; both endpoints and the count are the same under either family, since the widening moves no corpus by more than 0.33 points and moves neither extreme. Each of those numbers is a valid *lower* bound attained by a rule we can name and print (Proposition 1), so each licenses "at least this much of this corpus does not require the modality it is named for" as an arithmetic fact about the corpus and a named strategy.

As an *inference*, it does not follow that all five carry a signal, because a maximum over even five rules is a selection and inflates under a null where every rule is at chance. Tested against that selection (the confirmatory family-wise max-statistic permutation test of Table 2, over the frozen five only), 3 of 5 survive Holm. The class-level claim this paper makes is the second one, and it is the smaller one. The sentence to carry forward is 3 of 5 with a statistically supported blind-exploitability signal under a family fixed in advance, not five of five blind-solvable.

What the widened option family changed, per corpus. The option column of Table 1 is a maximum over all 9 rules, and the class-level condemnation below is stated twice so the amendment history is auditable. Under the five-rule *frozen* family fixed before any option corpus was read, every one of the 5 option-regime corpora already exceeds chance (the frozen maxima are 27.40%, 28.71%, 28.22%, 34.39% and 36.47% against chances of 20.00%, 25.00%, 25.00%, 33.33% and 33.33% respectively). The widened nine-rule family is a post-hoc sensitivity analysis on top: it moves the bound by 0.00, 0.17, 0.00, 0.00 and 0.33 points respectively, i.e. by at most 0.33 points on any corpus, and (the property that matters for inference rather than for the bound) it moves the significance verdict on 0 of 5 corpora (Table 2). The difference is not uniform. On EGOSCHEMA, VIDEO-MME and TEMPCOMPASS the widening changed nothing at all: the gain is 0.00 points and the winner is unchanged. On VSI-BENCH (choice) it moved the bound from 28.71% to 28.88%, a gain of 0.17 points, and the winner from OVERLAP to FIXED and QLENMATCH (2 rules attain that maximum exactly, so the single name in Table 1 is one of two and we print both here rather than let an enumeration order decide which is reported); on PERCEPTIONTEST from 36.47% to 36.80%, 0.33 points, and from OVERLAP to QLENMATCH. So the largest effect of doubling the family anywhere in this paper is 0.33 points, and the sentence to carry forward is that one, not EGOSCHEMA's zero. None of these numbers is itself the claim: a maximum over even the frozen five is a selection, and the family-wise test below decides which of the five excesses survive it: run twice, once over the frozen

family and once over the widened one, because a bound and a p-value do not tolerate a post-hoc family equally.

Which of these excesses are real, tested family-wise, and which family. An accuracy that is the maximum over an enumerated family is inflated by the selection even when every rule is chance-level, so each option-regime corpus gets a *family-wise max-statistic permutation test*: the gold is moved to a uniformly random option slot within each item, every rule is re-scored, and the statistic is the family maximum of the raw accuracy, so the null distribution *is* the distribution of the selection ($R = 2000$ replicates, seed 20260914). The five per-corpus p-values are then stepped down by Holm for the class-level statement (Table 2). We call it a max-statistic test rather than a max-t test because that is what it is: the statistic is the raw maximum accuracy, with no per-rule standardisation by a standard error, and a reader who saw "max-t" would be entitled to expect studentised statistics that are not there.

Which family the maximum runs over is a separate decision from which family the bound is taken over, and the two are made differently. Widening a family cannot invalidate a *bound*: a maximum over more named strategies is still attained by a named strategy (Proposition 1(ii)). It can invalidate a *p-value*, because four of the 13 scored strategies were chosen after two option corpora had been read, and a confirmatory test cannot be run over a family selected with the data in view. So the test is run twice on the same permutation replicates and reported in two halves. The **confirmatory** half maximises over the 5 rules frozen before any option corpus was read and is what every class-level statement in this paper reads. The **exploratory** half maximises over the expanded family and is a sensitivity analysis, with its own wider null: widening the family raises the observed maximum and the null's maximum together, which is why a larger β does not buy a smaller p.

The confirmatory verdict is 3 of 5: on EGOSCHEMA, TEMPCOMPASS and PERCEPTIONTEST the observed family maximum exceeds what the exchangeability null produces, by a rule we can print. The 2 that do not (VSI-BENCH (CHOICE) and VIDEO-MME) are part of the finding. The exploratory half agrees corpus for corpus (3 of 5, with the two verdict vectors differing on 0 corpora), so nothing in the paper's conclusions turns on the post-hoc widening; that agreement is worth reporting precisely because it was not guaranteed, and it is the reason the frozen family is what we quote rather than a restriction we had to accept. VIDEO-MME's positional rule (28.22% against 25.00%) sits inside selection noise ($p_{\text{Holm}} = 0.2468$). VSI-BENCH's headline excess dissolves entirely once its own option-count mix is priced into the null. Its items are not uniformly $K = 4$: the corpus mixes option counts, and because the permutation preserves each item's own count, a clean corpus of the same mix is what the observed family max is compared against. It loses. In 85.3% of the 2000 replicates the null's own family max reaches or exceeds the observed 28.71%, which is the $p_{\text{Holm}} = 0.8536$ of Table 2: the excess over the headline 25.00% is an artefact of the mixture and no rule in the frozen family clears the mixture-matched null. The expanded family does not rescue it either, and is further away ($p_{\text{Holm}} = 0.9780$), which is the widening's own null moving with it. The contrast with TEMPCOMPASS is the point. There the observed family max is reached by only 0.5% of replicates, so a Δ of $+1.06$ points that a plain z-test calls noise ($z = +0.88$) is significant against the null the corpus's own structure permits. A plain z-test against the headline chance (the obvious check, and the one a reviewer runs first) misses both, in opposite directions.

The corpus that was built not to leak, and does (barely). TEMPCOMPASS exists to require temporal structure, and its excess over the headline chance is 1.06 points, attained by LONG, the oldest artefact in the list. The naive reading is that 1.06 points on $n = 1579$ is noise ($z = +0.88$), and the headline chance understates how clean this corpus is: its option counts are mixed, and the permutation preserves each item's own count, so the null the observed maximum is compared against is a corpus of the same mixture rather than a uniform 33.33% one. Read against the intervention the corpus's structure actually permits, the length artefact is real, with only 0.5% of the 2000 replicates reaching the observed maximum: $p_{\max} = 0.0055$, $p_{\text{Holm}} = 0.0165$, under the frozen family. Small, positive, and significant is the sharpest reading available: the corpus's design worked, it did not work completely, and the instrument distinguishes those before a model is run.

The corpus we evaluate on ourselves. VSI-BENCH is one of the three witness columns the parent paper of this programme reports against, and the honest finding is that its option column does not survive correction under either family: the excess reported beside it in Table 1 is a property of the

corpus's option-count mix, not of a blind rule, and no rule clears a null that carries the same mix (Table 2). We would rather report that on a corpus we have an interest in defending than print an uncorrected maximum that indicts our own benchmark for a mixture artefact.

The largest bound in the paper, and why it is not the refinement trap. PERCEPTIONTEST reaches $\beta = 45.17\%$ against a null of 6.45%, an excess of 38.72 points, at granularity $g = 8$. §D.3 is the reason to distrust exactly that shape of number, so it has to be argued rather than reported. Two things separate it from the artefact. First, the excess is measured against the permutation null at the *same* granularity, so the part of β that small groups manufacture is subtracted rather than ignored. Second, and this is the reason it survives here and would not on OPENEQA, the corpus has 19140 items: at $g = 8$ the median group still holds 115 items, where on OPENEQA it holds one. The refinement trap is a statement about group size, not about g, and this corpus is nearly twelve times larger. What the number says is that a reader who sees only the first 8 tokens of a PERCEPTIONTEST question can name its answer text 45.17% of the time.

That figure also says something uncomfortable about the option-regime column beside it. The same corpus's option bound is 36.80%, and $45.17 > 36.80$: the free-form estimator, applied to the answer *text*, finds a regularity worth eight points more than anything the enumerated option family finds. A rule of the form "pick the option whose text is the template group's modal gold text" is blind, is enumerable, and would attain at least the free-form figure, so the option-regime family is understating this corpus by at least that margin. We report the gap rather than close it, because closing it means adding a rule to Definition 10 after seeing which corpus it helps, and because Proposition 1(ii) means the printed number remains a valid lower bound either way. It is the clearest available measure of how much a family of nine surface rules leaves on the table.

What the extra corpora broke, and it was the announcement count. V asks whether the gold answer appears verbatim in the question. On VSI-BENCH's option items that is true of 100.00% of them (every single one) because the stem enumerates the label space: *"is the table to my front-left, front-right, back-left, or back-right?"*. The gold is announced and so is every distractor, so the number carries no information and would have been the largest entry in the table. TEMPCOMPASS does the same thing differently, packing its options into the question field, where V on the corpus's own string is 100.00% against 0.51% on the stem alone.

Two corpora never exposed this, and the repair is not a special case. In the option regime the label space *is* the option set, so every gold is in it by construction and the question worth asking is whether the gold is announced **and the distractors are not**: that is what tells a blind reader which option to take. Writing V^{disc} for that count, VSI-BENCH gives $V^{\mathrm{disc}} = 0.00\%$ with 2.64 distractors announced per item on average: no leak at all, correctly. Definition 12 is amended accordingly, and the amendment is the kind that only arrives with a fourth corpus.

D.3 THE REFINEMENT CURVE, AND WHY A SINGLE NUMBER WILL NOT DO

Table 6 and Fig. 5 are Theorem 1 and Proposition 7 measured. At the corpus's own category labels the blind bound is 11.7% against a permutation null of 8.3%. One refinement step in, the estimate *rises* to 11.9% (this is the maximum of the sweep, not the coarsest value), while the median group has already fallen to a handful of items, and the excess over the null peaks at 5.3 points at granularity 4. By granularity 8 nearly every group holds one item and the estimate has collapsed to 1.6%.

Both failure directions are therefore available on one corpus by changing a choice that is usually left implicit. And granularity is not the only such choice: reporting the same estimator *in sample* rather than held out raises the figure to 97.2% at the finest granularity, because each item then votes for its own answer (Fig. 6). Taken together, **one corpus can be reported as 1.6% or 97.2% blind-solvable by varying two choices that a blind-baseline result usually does not state.** A blind-baseline number without its granularity and its held-out treatment is not a measurement, and the statistic that survives the sweep is the excess over the null rather than the estimate.

D.4 WHICH CATEGORIES ARE BLIND-SOLVABLE, NAMED BEFORE ANYTHING IS RUN

Two readings from Table 7 and Fig. 7. object state recognition has the largest announcement count at 28.2%, and a blind bound of 36.5% at g=1: on more than a third of that category, a strategy that

Table 6: The refinement curve on OpenEQA. g is the granularity of the question template: g=0 groups items by the corpus's own category label and each increment adds one leading question token. β is the leave-one-out modal accuracy inside a group, the null is the same estimator averaged over $R = 1000$ uniform permutations of the corpus's own answer multiset onto the same group-size profile (exact in that nothing about the answer distribution is modelled, Monte Carlo in that the average over re-deals is estimated rather than computed), and the excess is their difference. Refinement raises the excess, then destroys the estimate as groups collapse to one item: a blind-baseline number without its granularity is not a measurement.

g	groups	median n	β (%)	null (%)	excess
0	7	231	11.67	8.29	+3.39
1	92	2	11.86	7.05	+4.81
2	280	1	10.27	5.49	+4.78
3	586	1	7.95	3.42	+4.53
4	1008	1	6.72	1.39	+5.33
5	1320	1	3.79	0.56	+3.23
6	1478	1	2.26	0.27	+1.99
7	1533	1	2.14	0.14	+2.00
8	1571	1	1.65	0.09	+1.56

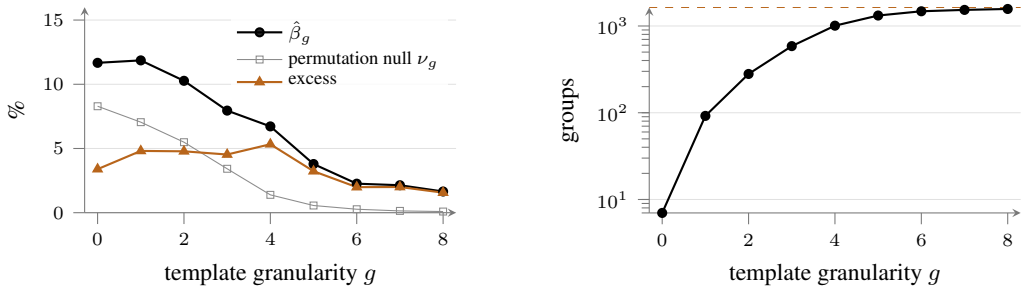

Figure 5: Theorem 1 and Proposition 7 on OpenEQA. Left: the estimate rises above its coarse value before collapsing, while the permutation null falls monotonically from 8.3% throughout, so the two move in opposite directions over the first refinement step; the excess is their difference and peaks at an intermediate granularity, which is how both directions of error come to be reachable on one corpus. The rise in $\hat{\beta}$ is a property of this corpus and this template family, not a consequence of the theorem; Table 1 lists the 4 rows on which it does not occur. Right: why the collapse must occur. The group count approaches one group per item, at which point the held-out remainder of every group is empty and the estimator is identically zero.

Table 7: Per-category counts on OpenEQA, ordered by excess. V is the share of items whose gold answer appears verbatim in the question. β is the blind bound at the coarsest granularity and $\beta_{g=1}$ one refinement step in, with the permutation null at g=1 beside it and the excess their difference. Both the null and the excess are reported at g=1 and not at g=0, because within a single category the coarsest partition is one group and re-dealing that group's own answers to itself makes the null equal the estimate by construction (Appendix G). The instrument names which categories are partly blind-solvable and by how much, before any system is run on them.

Category	n	V (%)	β (%)	$\beta_{g=1}$ (%)	null$_{g=1}$ (%)	excess
spatial understanding	220	5.5	13.64	16.36	9.96	+6.40
world knowledge	213	2.8	12.21	12.68	8.04	+4.64
attribute recognition	240	0.8	13.33	13.75	9.63	+4.12
object state recognition	252	28.2	36.11	36.51	34.53	+1.98
object recognition	231	0.0	2.60	2.60	1.56	+1.04
object localization	263	0.8	0.00	0.00	0.68	-0.68
functional reasoning	217	0.5	2.76	0.00	0.84	-0.84

never looks at a frame is correct. And spatial understanding carries the largest excess over the null at 6.4 points, on a bound of 16.4% one refinement step in (g=1). That is the category the literature already flagged as one where a text-only model is nearly as good as a multimodal one (Majumdar

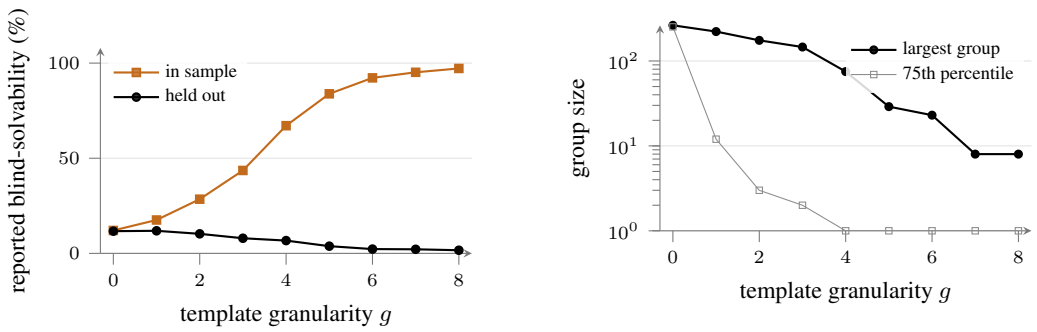

Figure 6: The two unstated choices, and the range they cover. Left: the same corpus reported with and without holding the item out. In sample the figure rises to 97.2% because each item votes for its own answer; held out it falls to 1.6%. *One corpus can therefore be reported as 1.6% or 97.2% blind-solvable by varying two choices that a blind-baseline result usually does not state.* Right: why, on a log scale. Beyond the first refinement step three quarters of groups hold a handful of items, and an in-sample mode inside such a group is a description of the group rather than a prediction about it.

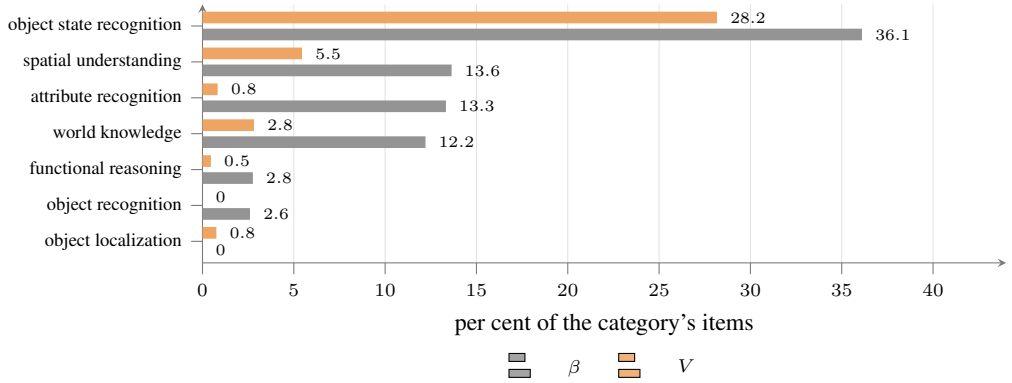

Figure 7: Per category on OpenEQA, ordered by the blind bound. Four of the 7 categories carry a β above ten per cent at the coarsest granularity, and the largest announcement count sits on the category with the largest bound. The instrument names them before any system is run on the corpus. The same numbers, with the permutation null and the excess, are in Table 7.

et al., 2024; Xu et al., 2026), and the instrument reaches it from annotations alone, with no model and no frames.

Fig. 9 puts the two original corpora side by side against their own baselines, which is the comparison the two answer regimes do permit: each against the value its own estimator returns in the absence of structure, and neither against the other.

D.5 A positive control, on a corpus whose defect was known first

Every other number in this paper is a claim about a corpus we cannot check independently. This one is not. The parent corpus (Anonymous authors, 2026c) shipped in two versions differing only by a documented defect (§C.6), and the instrument is run on both under the settings used everywhere else; Table 3 is the result. A second control, on a repair performed and documented by somebody else (NLVR2's balanced split), is reported in Appendix G, and is the arm's-length calibration this one cannot be.

On the 8 order-comparison items the instrument reports $\hat{\beta} = 100.00\%$ before the repair and 0.00% after it. The first is the maximum the statistic can take and the second is its minimum, on the same questions about the same frames, differing only in which of two objects the generator named first. Diluted into the whole corpus the signal survives but shrinks (17.50% against 7.50%, a drop of 10.00 points, and an excess over the corpus-wide permutation null of 12.13 against 3.14 points), which is a

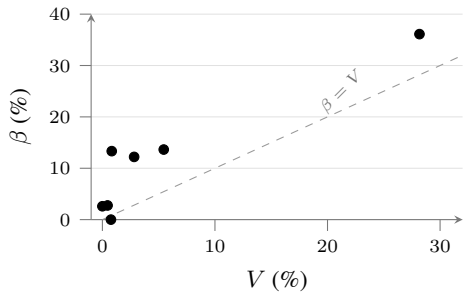

Figure 8: The same categories in the (V, β) plane. 6 of the 7 lie above the diagonal and one (OBJECT LOCALIZATION, at $V = 0.76\%$ against $\beta = 0.00\%$) lies below it. That is a *measurement* on this corpus and not Proposition 2, which orders the two counts as detectors of one failure and explicitly declines to order their rates; the single point below the diagonal is the proposition's refusal showing up inside one corpus, as VSI-BENCH (choice) in Table 1 shows it across corpora with V far above β. Points well above the diagonal are categories whose blind-solvability comes from somewhere other than verbatim leakage.

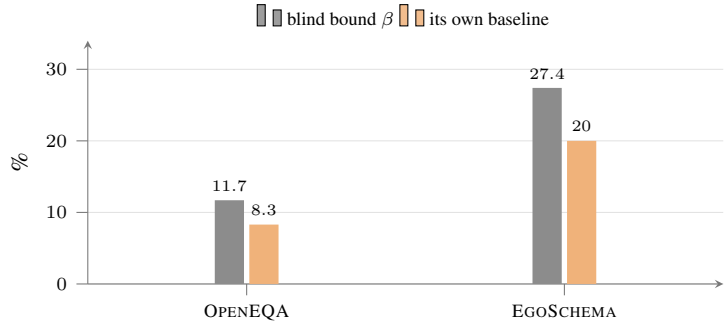

Figure 9: Two corpora, two answer regimes, one instrument. The baseline is the permutation null for the open-answer corpus and chance for the option corpus, because those are the values each regime's estimator returns in the absence of structure. Both corpora exceed their baseline, and the excess is what the instrument certifies; neither number is comparable across the two regimes and we do not compare them.

useful calibration in itself: a defect confined to a tenth of a corpus moves the corpus-level number by a few points, so a corpus-level β should be read as a floor on a possibly severe local defect rather than as a diffuse property.

And one qualification to our own recommendation. §2 says the statistic to report is the excess over the null rather than β alone, and on the eight order items the excess is 0.00 before the repair and 0.00 after: it does not move, because a within-group null re-deals that group's own answers onto that same group, and eight identical golds re-deal to eight identical golds. A constant gold is invisible to a within-group null by construction. It is not invisible to β, which is at its maximum, nor to the corpus-wide null in the row above, where the other subsets' answers enter the shuffle. The recommendation therefore needs its scope stated: report the excess, and read β alongside it, because the two fail on different defects.

Two details deserve emphasis because they are what make this a control.

The instrument had no opportunity to be tuned to it. The defect was found by hand, its mechanism was read off the generator's source, and it was written into the parent paper's deviation log before this instrument was written. The direction of the expected result was fixed in advance, and the instrument has no free parameter with which to arrange it.

V does not move. The announcement count is 15.00% in both versions. The defect is invisible to it, because a constant gold is not an announced gold: nothing in the question states the answer. This is Proposition 2 seen from the other side: that announcement is a detector of the same kind of failure does not make V a weak version of β, it makes them detectors of different defects, and §4.6 measures how different.

An honest reading of 0.00%. The post-repair value is 0, and chance on a balanced binary item is 50%. That is not a paradox and it is not a bug; it is Lemma 1 at its smallest scale. With four BEFORE and four AFTER, holding out any item leaves the opposite label in the majority, so the leave-one-out modal rule is wrong on every item by construction. The estimator is one-sided, and its numerical value is not an accuracy, a fact that this paper argues about granularity in general and that the control exhibits at $n = 8$. A reader who wanted an accuracy here should read the permutation null, which Table 3 now prints beside every β in it: after the repair the null on those eight items is 0.00% too, which is the honest statement that the estimator has nothing to report there rather than that the corpus is at chance.

D.6 TWO LEAKS AND AN OVERFIT, MEASURED

Proposition 3 claims one mechanism behind three pathologies. Here it is, on OPENEQA, across the refinement axis: Table 4 for the estimators, Table 8 for the bootstrap, and Fig. 3 for both.

The in-sample leak. Scoring the modal rule without holding the item out gives 12.04% at $g{=}0$, where every group is large, and 97.19% at $g{=}8$, where 1523 of 1571 groups contain a single item and each item is its own mode. Nothing about the corpus changed between those two numbers. This is the leak the field knows about, and it is still what a blind baseline computed without an explicit held-out split silently reports.

The small-group overfit. Holding the item out removes the leak entirely (the answer is not in the evidence at any group size) and replaces it with a mode fitted to almost nothing. The reported curve moves 5.14 points across $g \in \{0, 1, 2, 4\}$, and 10.21 points across the whole sweep, and it does so non-monotonically, rising before it collapses. That shape is what Proposition 7 says is available and what Table 1 says 4 of the 9 rows do not exhibit; Theorem 1 guarantees only the collapse.

The permutation null moves in the opposite direction over the same range, from 8.29% to 0.09%, and Lemma 1 says why; but the lemma's bounds are on $\nu(n)$ for a *group* of size n, and the reported null is an average over groups, so the two must not be conflated. Per group, $\nu(n)$ runs from 0 at $n = 1$ through the collision probability 1.54% at $n = 2$ up toward the base rate 8.80% as n grows. The corpus-level average therefore starts near the base rate at $g{=}0$, where groups are large, falls through the collision probability between $g{=}4$ and $g{=}5$, and continues *below* it to 0.09% once singletons with $\nu(1) = 0$ dominate the average. Both endpoints of the lemma are computed from the answer distribution alone; neither is a floor on the average, and the appendix works through why the measured value ends up under $\sum_y p_y^2$.

The bootstrap leak. This is the one we did not anticipate. The ordinary nonparametric bootstrap gives a mean of 46.74% with a 95% interval of $[43.40, 49.57]$ at $g{=}4$, against a point estimate of 6.72%, an inflation of $6.96\times$. The inflation grows with refinement ($0.99\times$, $1.32\times$, $2.26\times$, $6.96\times$ at $g = 0, 1, 2, 4$), which is what Eq. (1) predicts: a duplicate is only decisive when the group is small enough for one vote to carry the mode.

We diagnose rather than infer. Each draw is scored twice, once as drawn and once with duplicates removed from that same draw and nothing else altered. The deduplicated column reads 11.02, 10.57, 10.56 and 5.34 against point estimates of 11.67, 11.86, 10.27 and 6.72: the inflation is gone at every granularity, and what is left is a residue of at most 1.38 points. That residue is expected and is not noise. Removing duplicates from a draw of n from n leaves about 63% of the corpus, whose template groups are correspondingly smaller, and a leave-one-out estimate on smaller groups is lower, which is the same Theorem 1 effect measured a third way. The deduplicated value therefore lands slightly *below* the point estimate in 3 of the 4 rows, as it should; in the remaining row it sits 0.29 points above, which is smaller than every below-gap and is the size of the residual noise the diagnosis is read against. The measured duplicate rate is 0.418 against the asymptotic 0.418 of Eq. (1).

The practical consequence is a recommendation we would rather not have had to make: do not put a bootstrap interval on a leave-one-out blind bound. The permutation null is the right object, it is exact for the observed answer multiset, and it resamples nothing.

How far this reaches beyond blind bounds. Nothing in the mechanism is about blindness, about annotations, or about this instrument. The ingredients are three, and they are generic: a statistic fitted

Table 8: Why β must not be quoted with a bootstrap interval. Resampling items with replacement puts duplicates of an item into its own template group, and the leave-one-out rule then scores the held-out item against a surviving copy of itself. The last column removes duplicates from the *same* draw, changing nothing else, and the inflation disappears: the cause is therefore the leak and not the resampling. At $g{=}4$ the interval sits at $[43.40, 49.57]$ around a point estimate of 6.72: a paper quoting it would be reporting an artefact of its own error bars, and reporting it as rigour. The permutation null already in the instrument is exact for the observed answer multiset and resamples no items at all.

g	point β	bootstrap mean	95% interval	inflation	same draw, deduplicated
0	11.67	**11.51**	$[8.86, 13.51]$	$0.99\times$	11.02
1	11.86	**15.60**	$[12.65, 17.85]$	$1.32\times$	10.57
2	10.27	**23.25**	$[19.99, 26.10]$	$2.26\times$	10.56
4	6.72	**46.74**	$[43.40, 49.57]$	$6.96\times$	5.34

within groups, scored on the same units it was fitted on with those units held out, and resampled at the unit level. Resampling at the unit level returns duplicates of a held-out unit to its own group, which is the one thing the held-out design exists to prevent, so the interval is computed on draws in which the estimator's defining property has been silently removed. Any transductive or group-conditioned evaluation with that shape is exposed: a per-cluster or per-user mode or mean scored leave-one-out, a nearest-neighbour or retrieval baseline scored inside its own index, a per-template or per-slice statistic with resampled error bars. We measure it only where we found it, and we make no claim about the size of the effect elsewhere: it is governed by how small the groups are, which is why it grows with granularity here. But the diagnosis transfers whole, and it is cheap: rescore the same draw with duplicates removed and change nothing else. If the interval moves, the interval was measuring the duplicates. We state this because a reader whose estimator has that shape should run the check, and because the correct scope of the finding is wider than the corpora it was found on, while the corpora are all it was tested on.

And the repair. The split estimator reads 11.20, 11.05, 11.89 and 11.78 at $g = 0, 1, 2, 4$, a range of 0.84 points where leave-one-out over the whole corpus moves 5.14.

That comparison is not like for like, and the honest one is smaller. Leave-one-out is scored on every item, including the items of singleton groups, whose held-out remainder is empty and which therefore score zero by construction; the split cannot score those groups at all. Restricted to the groups the split can see, leave-one-out reads 11.67, 12.09, 11.44 and 13.35: a range of 1.91 points, a little over twice the split's 0.84, and a little over a third of the 5.14 the previous comparison suggested. Most of the apparent collapse of leave-one-out is the singletons entering the average, and the part that is the estimator rather than the population is the smaller number. The refinement trap is therefore substantially, not entirely, a property of *how* the baseline is scored rather than of the quantity being scored.

The caveat is coverage, and it is severe: 100.0% of groups are scorable at $g{=}0$, 19.4% at $g{=}4$, and 3.1% at $g{=}8$. By items (the unit the estimators are actually weighted in) that is 100.0%, 50.4% and 6.9%. The apparent rise to 28.67% in the last row is a statement about the 6.9% of items in groups that still contain two, which are not a uniform sample of anything, and we do not read it.

D.7 Does the mask predict what a blind system actually solves?

§4.2 validates the instrument against a documented defect, which establishes that it fires when it should. It does not establish that what it flags is what a blind system gets. That is a claim about behaviour, and the instrument's whole design is to avoid running anything; testing it therefore requires stepping outside the instrument for the only time in this paper. The masks below inherit the single-valued rule: the blind mask is the hit set of one named blind strategy, so the comparison tests a strategy, not an oracle.

We run 5 open backbones on both corpora with *no frame, video or image in any request*. The models are blind by construction rather than by ablation: there is nothing to ignore. Scoring is exact on EgoSchema, which is multiple choice, and by normalised token containment against the gold on OpenEQA. Containment errs in both directions and we do not claim otherwise: it under-counts

paraphrases, and it over-counts, since a prediction containing the gold's tokens inside a negation ("not brown" against gold "brown") scores correct. Both errors fall on the flagged and unflagged arms alike, so neither can manufacture a difference between them, which is the only property the test needs. 12 of the 1636 OPENEQA items are not scored at all, because their golds (*on*, *off*) consist entirely of stopwords and so have no content tokens to contain; they are excluded rather than scored either way, which is the same rule the announcement count uses, and $n_{\text{in}} + n_{\text{out}} = 1624$ in Table 5 accordingly. No judge is used anywhere.

The masks are the instrument's own outputs and their granularity is part of them. On OPENEQA the mask is the leave-one-out hit set of the modal rule at $g{=}1$: it is therefore derived from the other items of a template group and is a function of the corpus, not a fixed strategy in the sense of Definition 1, a distinction that matters because $g{=}1$ is where §4.3 says the small-group overfit begins. On EGOSCHEMA the mask is the item set the OVERLAP rule settles, which is a fixed strategy and reads no answers.

That the OPENEQA mask is read at the granularity this paper itself calls unstable is a fair objection to raise before the table rather than after it, and it has a direct answer, because the objection is about the partition and not about the runs. The stored predictions are per item, so the mask can be re-derived at another granularity and the same predictions re-split at no cost. Appendix G does that at $g{=}0$, $g{=}1$ and $g{=}2$ (partitions differing by more than an order of magnitude in group size, 231 items at $g{=}0$ against 1 at $g{=}2$), and all 15 of the 15 resulting cells separate the arms in the same direction, 15 of them surviving a Holm step-down over the whole sweep, at lifts of 1.56–$5.44\times$. The separation reported below is therefore not manufactured by the small groups $g{=}1$ produces. It remains true that *which* items are flagged depends on g, which is Theorem 1 rather than a defect and is why the released artefact is a mask with its granularity attached.

The comparison is within a corpus and within a model, which is what makes it a test of the instrument rather than of either. Comparing a model's blind accuracy against β as two scalars would prove nothing: both are accuracies of blind procedures and need not be ordered either way.

On the open-answer corpus the mask is strongly predictive. On OPENEQA, every one of the 5 backbones solves the flagged items at a higher rate than the rest, by 13.7–46.1 points, a lift of 2.31–$4.64\times$, and all 5 differences are significant after a Holm correction over the 10-test family of Table 5. Accuracy inside the mask runs 24.2–58.8% against 9.3–13.5% outside it, on models whose overall blind accuracy is only 11.8–18.2%. The announcement mask, which §4.6 shows flags almost disjoint items, is independently predictive by 13.4–35.2 points on 5 of 5 backbones. Two detectors built from annotations alone, each identifying a different population that a blind model disproportionately solves.

On the option corpus it is predictive for weak blind models and not for strong ones. On EGOSCHEMA the difference goes the right way on 3 of 5 backbones, but only 1 of them survives a Holm correction over the 10-test family, by 14.0 points. Of the rest, one separates before correction and not after, one is positive but never separable, and two are reversals, the larger -6.8 points. Counting every positive difference as a separation, as an earlier version of this section did, treats 5.7 points on 137 and 363 items as evidence (a gap this corpus has no power to resolve), and a $p = 0.047$ as confirmation when it is one of 10 tests; it is not. That is a negative result and it is the one we consider most informative in the paper, because the instrument's own theory predicts it.

Proposition 1 says $\hat{\beta}$ is a maximum over an *enumerated* family and that acquittal is unavailable because the family cannot be completed. The mask is therefore the item set solved by a named strategy, not the item set a blind system will get. A weak blind model falls back on exactly the surface regularity the family names, so the two coincide; a capable blind model reaches items by world knowledge and plausibility, which is a strategy outside the family, so they come apart. The measurement makes that concrete: the best enumerated rule attains 27.4% on EGOSCHEMA, and a blind backbone attains 47.4% on the same items, *above* the instrument's bound by a wide margin, in the direction one-sidedness says is the only one available.

We take three things from this. The instrument is validated where its estimator matches the strategy in play, and it is a bound and not an estimate everywhere else, which is what it claims to be. A blind backbone scoring 47.4% on a five-way video benchmark, with chance at 20%, is a sharper condemnation of that corpus than any annotation-side number we can compute, and it costs a model

Table 9: The model-side intervention on OPENEQA: the same backbones, the same mask, the same scorer, and one stated synonym substituted for each question's leading token. The mask is the leave-one-out modal hit set at $g{=}1$ (the corpus's category label plus that leading token), so the perturbation moves the surface feature the mask is partly built from while leaving every content token alone. It is held fixed at its original value: re-deriving it on perturbed questions would test a different mask. GAP is IN minus OUT within a run, and DID is the baseline gap minus the perturbed gap. DID (strat.) is the same contrast computed within each perturbation-severity stratum and pooled by stratum size, which is the headline because the severity is correlated with the mask; its interval is the standard error of the paired per-item difference, the two runs having scored the same items. A positive DID is the separation shrinking when the interrogative moves, which is what a mask reading the leading token rather than the question's content would produce. p_{Holm} steps down across the 2 backbones. No frame, video or image was sent in any call, here or in the baseline. No stratified estimate is distinguishable from zero after the step-down ($|z| \leq 1.03$), and the two carry opposite signs, so the separation does not close when the interrogative moves; §4.5 gives the effect size this rules out and the one it cannot.

Backbone	IN	OUT	gap	IN′	OUT′	gap′	DiD	DiD (strat.)	p_{Holm}
Qwen3-VL-8B	50.5	13.5	+37.0	47.9	13.7	+34.2	+2.8	**+3.6** $\pm$ 3.5	0.604
Qwen3.8-27B	58.8	12.7	+46.1	61.9	12.2	+49.6	-3.5	**-1.0** $\pm$ 3.1	0.748

run. And the gap between the two is the practical measure of how much a bound of this kind leaves on the table: on this corpus, about twenty points.

What this does not license. It does not turn $\hat{\beta}$ into a predictor of a particular system's accuracy, and Proposition 10 still forbids subtracting one from the other. What it supports is narrower: the item sets the instrument returns are not artefacts of its estimator, because a system that never saw them behaves differently on them. That is the property Definition 11 needs, and it is the reason we recommend reporting conditional accuracy on the unflagged subset rather than a difference of scalars.

D.8 AN INTERVENTION ON WHAT THE MASK IS BUILT FROM

§4.4 is a between-item comparison and therefore a correlation, which the limitations have said since the first version: an unobserved third property could make an item both flaggable and easy. The test that separates those is a within-item one, and the OPENEQA mask makes it available, because the mask is built out of a surface feature we can move. It is the leave-one-out modal hit set at $g{=}1$, and $g{=}1$ is the corpus's category label plus the question's *leading token*. So we substitute one stated synonym for that token (`what` → `which`, `is` → `do you observe that`, and a fixed map given in `scripts/a2_interventions.py`), leave every content token alone, re-ask the same backbones with no frame in any request, and score against the same gold with the same scorer, which is checked by recomputing both runs' stored scores rather than assumed. The mask is held fixed at its original value: re-deriving it on perturbed questions would test a different mask.

The estimand is the difference in differences: the in/out gap under the original questions minus the gap under the perturbed ones. A large positive value is the separation collapsing when the interrogative moves, which is what a mask reading the leading token rather than the question's content would produce; a value near zero is the separation surviving a perturbation of the feature the mask is partly built from.

The naive version of that estimand is confounded, and we report the repaired one. The synonym map is not uniformly severe: some leading tokens are replaced by one word (`what` → `which`) and others by a phrase (`is` → `do you observe that`). That severity is correlated with the mask, and not weakly: items receiving a multi-word rewrite are inside the blind mask 18.4% of the time against 5.5% for a one-word swap. The reason is mechanical and, in hindsight, obvious: the flagged items are the ones whose answers are modal, which on this corpus are largely the yes/no questions, which are exactly the ones beginning `is`, `are`, `does` or `do`. A difference in differences computed over the whole corpus therefore mixes the effect we want (the cue being removed) with an effect we do not: the flagged arm being asked a clumsier question. We found this before reading the result rather than after, which is the only reason it is reported as a design property instead of a robustness check.

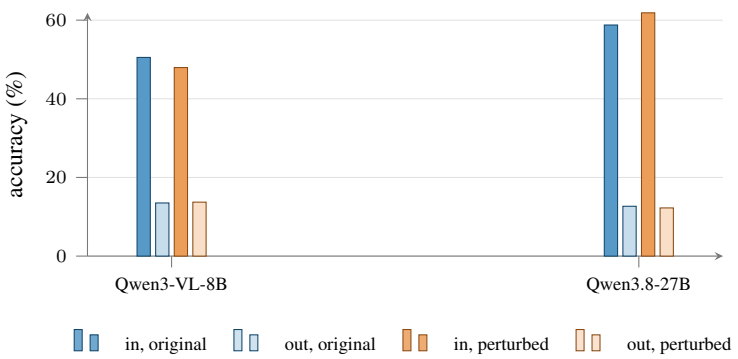

Figure 10: The leading-token intervention on OPENEQA. Each backbone is shown four times: inside and outside the instrument's $g{=}1$ blind mask, under the corpus's own questions and under questions whose leading token has been swapped for a stated synonym with every content token left alone. The mask is held fixed at its original value in both arms. The quantity to read is not any single bar but the change in the *gap* between the dark and light bars of a pair, which Table 9 reports as a difference in differences. That gap is 37.0–46.1 points under the original questions and 34.2–49.6 under the perturbed ones: it does not close. No frame, video or image was sent in any of these calls.

The repair is to compute the contrast *within* each severity stratum and pool by stratum size, which removes the composition effect entirely; Table 9 reports that stratified estimate as the headline and the unstratified one beside it, so the size of the confound is visible. The perturbation's own difficulty cost is measurable separately, on the unflagged arm where the mask plays no part, and it is what the two strata differ by.

What stratification does *not* buy is a homogeneous treatment, and the estimand should be named accordingly. A one-word synonym substitution and a multi-word syntactic rewrite are two different interventions applied to two different sets of items, so this design is a *heterogeneous perturbation test* whose estimate is an average over a treatment that varies across items, not a single treatment with a constant dose. Stratifying by severity removes the part of that heterogeneity which is correlated with the mask (the part that would bias the contrast), and leaves the part that makes the pooled number an average of two different things. So we ran the design that does not need the stratification, and it is the next subsection. This one is reported first and in full because it is what was run first, and because the difference between the two is itself a measurement of what the confound was worth.

Two further properties of the estimator, both fixed before the result was read. The variance is computed on the *paired* per-item difference rather than from four independent binomials: the two runs score the same items, so the in-arm's two proportions are positively correlated, and treating them as independent adds a covariance that is really there. And the mask arm is small (194 flagged items against the rest), so the study is powered to detect a large collapse and not a small one: at the standard errors actually achieved, the smallest effect this design could have resolved at conventional power is 11 points pooled and 16 points inside the one-word stratum, against baseline gaps of 13.7–46.1 points; the test can therefore rule out the leading token carrying more than roughly 24–29% of the separation, and cannot rule out less. A null result here is therefore evidence against a *large* leading-token effect and is not evidence of no effect, which is the reading we take and the one §5 repeats.

The result: the separation survives the perturbation. The leading token actually moved on 92.1% of the aligned items, so the intervention fired on nearly all of them. On neither backbone does the gap collapse. It is 37.0–46.1 points before the substitution and 34.2–49.6 points after it, and the stratified difference in differences is -1.0 and +3.6 points: neither distinguishable from zero ($|z| \leq 1.03$), *with opposite signs*, and neither surviving the step-down across the 2 backbones ($p_{\text{Holm}} \geq 0.604$). 0 of 2 are significant. The opposition of the signs is itself informative: a real cue being removed would push both backbones the same way, and what we see instead is the scatter of two null estimates. The naive and stratified estimates differ by at most 2.5 points, which is the size of the severity confound that the stratification removes and the reason it is the stratification, not the naive figure, that is read.

The conclusion we draw is bounded, and deliberately so. The blind advantage on OPENEQA is not carried, to the precision this design can reach, by the interrogative the mask is partly built from;

Table 10: The *matched-dose* intervention on OPENEQA: the same backbones, the same fixed mask, the same scorer and the same baseline runs as Table 9, but every leading token is replaced by exactly one token: a polar auxiliary by the same auxiliary in the other tense (`is` → `was`, `can` → `could`), a wh-word by another wh-word (`what` → `which`, `where` → `whereabouts`). Two defects of the earlier map are removed by construction. The dose is constant: the word-count difference between every question and its perturbation is exactly zero, checked against the strings that were sent, so perturbation severity is orthogonal to mask membership and there is nothing left to stratify. And the perturbed question stays *grammatical*: a leading-token substitution does not undo subject–auxiliary inversion, so the earlier map turned *"Is the front door open?"* into *"do you observe that the front door open?"*, an ungrammatical string, and one the flagged arm received far more often than the unflagged arm, since polar questions are most of what the mask flags. Items whose leading token has no one-token grammatical equivalent (`how`, `why`, `when`, and the non-interrogative openings) are *not* treated and are excluded rather than pooled: 1432 items are treated, 191 of them inside the mask, and the leading token moves on 100% of them. DID is the baseline gap minus the perturbed gap, with the standard error of the paired per-item difference; p_{Holm} steps down across the 2 backbones, separately from Table 9's family. No frame, video or image was sent in any call.

Backbone	IN	OUT	gap	IN′	OUT′	gap′	DiD	p_{Holm}
Qwen3-VL-8B	49.7	14.4	+35.3	46.1	14.9	+31.2	**+4.2** ± 3.0	0.324
Qwen3.8-27B	58.6	13.5	+45.1	57.6	13.6	+44.0	**+1.1** ± 2.5	0.656

it survives having that token replaced while every content token is left alone. That is a genuine strengthening of §4.4: the correlation there is not explained away by the most obvious lexical artefact available, which is the one the mask itself reads. It is not a demonstration that no surface feature carries it, and §5 states the residual. We note, without leaning on it, that the one direction in which both backbones agree is the one-word stratum, where the gap narrows by +5.1 and +5.3 points, an effect well inside noise at these arm sizes, which we report because it is the only coherent signal in the table and not because we can distinguish it from zero. It is worth holding onto for one paragraph, because the one-word stratum is the part of this design whose dose is already matched, and the next subsection makes the whole experiment look like it.

D.9 THE SAME INTERVENTION WITH THE DOSE MATCHED

The design above has two defects that are really one defect, and both are repairable without touching the mask, the scorer or the baseline runs. The first is the heterogeneous dose just described. The second we found while repairing the first, and it is the worse of the two: a leading-token substitution does not undo subject–auxiliary inversion, so *"Is the front door open?"* was asked as *"do you observe that the front door open?"*, which is not a grammatical English sentence. That is a difficulty shock rather than a cue removal, and it did not fall evenly: polar questions are most of what the mask flags, so the ungrammatical arm was disproportionately the arm under test.

The repair is a map in which every leading token is replaced by exactly *one* token of the same syntactic category, chosen so the sentence stays grammatical. A polar auxiliary becomes the same auxiliary in the other tense (`is` → `was`, `are` → `were`, `can` → `could`, `does` → `did`); a wh-word becomes another wh-word (`what` → `which`, `where` → `whereabouts`). Inversion is preserved because nothing is added or removed, so *"Was the front door open?"* is grammatical, every content token is untouched, and the leading token (the thing the $g{=}1$ template is built from) still moves. The tense shift is the one semantic change, and it is appropriate rather than merely tolerable here: OPENEQA asks about a recorded scan, so the state being asked after is in fact past. Leading tokens with no one-token grammatical equivalent (`how`, `why`, `when`, and the non-interrogative openings) are not treated at all and are excluded rather than pooled, since an untreated item carries no treatment whose effect could be averaged.

Three properties of this arm are checked against the strings that were actually sent rather than against the map's intent. The dose is constant: the word-count difference between every question and its perturbation is exactly zero, so perturbation severity is orthogonal to mask membership *by construction* and there is nothing left to stratify: a mismatch is a hard error in `analysis/a2_interv.py` rather than a caveat. The intervention fires on 100% of treated items, against 92.1% for the earlier map. And coverage of the arm under test is nearly complete: 1432 items are treated, 191 of them inside the mask, which is all but three of the 194 the mask flags.

The result is the same, and it is now cleanly estimated. The gap goes from 35.3–45.1 points to 31.2–44.0, and the difference in differences is +1.1 and +4.2 points, neither distinguishable from zero ($|z| \leq 1.40$, $p_{\text{Holm}} \geq 0.324$); 0 of 2 are significant. The separation survives a perturbation of the feature the mask is partly built from, under a treatment that is the same on every item it is applied to.

Two things did change, and both are improvements in the estimate rather than in the finding. The two backbones now agree in sign: 2 of 2 positive, where the heterogeneous arm produced -1.0 and +3.6 with *opposite* signs. We had read that opposition as the scatter of two null estimates, and that reading survives: what the matched dose removes is the part of the scatter the unequal treatment was manufacturing. And the bound tightens. At the standard errors this design achieves, the smallest difference in differences it could have resolved at conventional power is 9 points, so a null here rules out the leading token carrying more than roughly 20–26% of the separation, against 24–29% from the heterogeneous design. That is the honest content of both nulls, and the matched one is the stronger of them.

What neither design can do is unchanged, and no amount of dose-matching reaches it. This perturbs one surface feature, so it speaks to the leading token and to nothing else; a null is evidence against a *large* lexical effect and is not evidence of no effect. The conclusion is therefore the same one, drawn from a better experiment: the blind advantage on OPENEQA is not carried, to the precision two designs can now reach, by the interrogative the mask is partly built from.

This points the same way as the tie-break decomposition of §4.1, and the two are worth reading together even though they measure different things: a bound's excess there, a behavioural gap here, on different corpora and different regimes. On EGOSCHEMA the larger half of the excess turned out to be positional rather than lexical; on OPENEQA moving the question's lexical surface does not move the gap by an amount we can resolve. Neither result is about the questions' words, and that is the unified reading: what these corpora leak blindly is a property of how their answers are *distributed* across groups and slots, not of how their questions are *worded*. It is also why equalising or paraphrasing question text is not the repair a benchmark author should reach for: the release-side intervention that does address the larger half (shuffling the gold's slot) acts on the answer layout, which is the point §D.11 reaches from the other direction.

D.10 DO THE TWO COUNTS FIND THE SAME ITEMS?

Proposition 2 makes announcement a detector of the same kind of blind failure, which invites reading it as a weak version of the blind bound. On OPENEQA they behave as almost disjoint detectors.

Of 1,636 items, 94 are announced and 194 are predictable from the modal rule at $g{=}1$. If the two were independent, 11.1 items would be both. Measured, 3 are (a lift of 0.27, that is, the two *avoid* each other), and together they implicate 17.42% of the corpus. Proposition 2 is about the *kind* of failure the two counts detect, not about which items they land on: announcement narrows the answer to a candidate set the question fixes, the modal rule guesses from the question's form, and on this corpus the items where the answer is stated are largely not the items whose answer is guessable.

The reading we take from this is procedural. Reporting only β would miss most of the announced items; reporting only V would miss most of the blind-solvable ones; and reporting their sum would be wrong, because by Proposition 2 the two are counts of one failure reached by two routes. What a corpus audit should report is the union, with the intersection stated, and that is what §5 recommends.

D.11 THE OPTION REGIME: A DEFENDED ARTEFACT AND AN UNDEFENDED ONE

On EGOSCHEMA the modal-answer estimator returns 0.8%, which is far below chance and is an artefact of applying the open-answer instrument to an option corpus: an option's full text is nearly unique per item, so there is no mode. In the option regime the bound is 27.4% against chance 20%, an excess of 7.4 points, and the rule that achieves it reads only the token overlap between the question and each option (Table 11, Fig. 11).

The rules that fail are the informative part. The longest-option rule reaches 13.6%, *below* chance, which is what a corpus looks like when its authors equalised option lengths on purpose. The corpus was defended against the artefact its authors knew about; the overlap artefact was not defended, and it is worth 7.4 points to a system that never watches the video.

Table 11: The option-regime bound on EGOSCHEMA, 5-way with chance 20%. Each row is a blind strategy stated in advance and evaluated exactly; the bound is their maximum, which keeps it one-sided. The length artefact is *below* chance, so the corpus defended against the one its authors knew about; question–option overlap was not defended and carries 7.4 points.

Blind strategy	accuracy (%)
most question–option token overlap	27.40
modal gold index	23.40
shortest option	23.00
longest option	13.60
most tokens unique to the option	12.60
chance	20.00

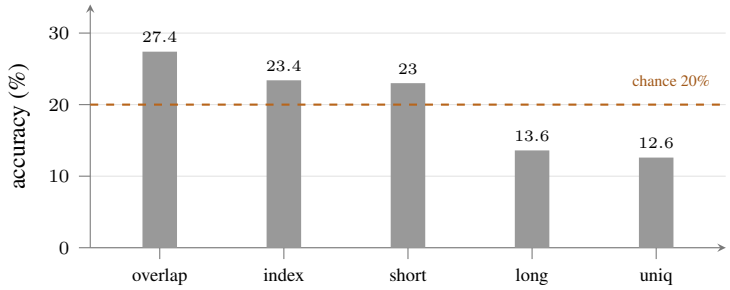

Figure 11: The option-regime bound on EGOSCHEMA. Five blind strategies, fixed before looking and reported in full including the ones that fail. Question–option token overlap carries 7.4 points over chance while the longest-option rule sits below it: the corpus was defended against the length artefact and not against this one. Enumerating the family in advance is what keeps the maximum a bound rather than a selection effect.

How much of that is overlap, and how much is position. The modal-index rule reaches 23.4%, which invites the conclusion that position carries little on this corpus. The conclusion is wrong, and the audit of Appendix G is what shows it: position is not carried by the rule named for it but by the tie-break inside the rule that wins. OVERLAP scores each option by its content-token overlap with the question, and on 83.0% of items that score does not separate the options at all; on 60.8% of them every option ties, usually at an overlap of one, because the question names the camera wearer and so does every option. On a tied item it is the family's index ordering that picks.

The two parts separate exactly, because an item's top score is either unique or it is not, and the buckets' hit counts sum to the rule's own accuracy. Of OVERLAP's 27.40% against chance 20.00%, the items where the score genuinely decides are 17.0% of the corpus and contribute +1.80 points of the excess; the items where the tie-break decides are 83.0% and contribute +5.60. On the fully tied items the rule *is* a constant-index rule, and it scores 24.34% there against a chance of 20.00%.

So the finding on EGOSCHEMA survives and its attribution changes: the corpus is blind-solvable above chance by a rule we can print, and the larger half of that is positional rather than lexical. Both halves are blind, both are in the enumerated family, and neither needs the video, which is why the bound and the condemnation are unaffected (§5, item 3b). What changes is the advice a benchmark author takes from it: equalising option lengths, which these authors did, does not address either half, and shuffling the gold's slot at release addresses the larger one.

Widening the family cannot lower the bound, and on this corpus it does not raise it either. Definition 10 began at five rules fixed in advance. Because $\hat{\beta}$ is a maximum over an enumerated family, adding rules can only raise it (Proposition 1(ii)), so a natural worry is that the family is small enough to be understating the corpus badly. We tested that by adding four more, each an artefact reported somewhere in the multiple-choice literature and none requiring the modality: *centroid*, the option most similar to the others, which fires when distractors are written by editing a correct answer; *outlier*, its complement; *fixed*, the single best constant index; and *qlenmatch*, the option whose length is closest to the median.

Over all 9 rules on EGOSCHEMA the bound stays at 27.40% and the winner stays OVERLAP: the gain from doubling the family is 0.00 points (Table 12). That is corpus-specific, and §4.1 gives the rest: the gain is also zero on VIDEO-MME and TEMPCOMPASS, and it is 0.17 and 0.33 points on VSI-BENCH (choice) and PERCEPTIONTEST, where an added rule wins outright. So the correct statement is that doubling the family moved the bound by at most 0.33 points on any corpus here. That is the only evidence available that the enumerated family is not leaving a great deal on the table, and §4.1 gives one measured reason to think it does leave something: on PERCEPTIONTEST the free-form estimator finds 45.17% where the option family finds 36.80%. It remains evidence and not proof: Proposition 1 is not repealed by a family that grew a little.

Two of the added rules are informative even though they lost. *fixed* reaches 23.40% at index 4, above the chance level of 20.00%: the gold sits in one option slot more often than it should, a positional bias of the kind that is trivially exploitable and trivially fixed by shuffling at release. And *centroid* reaches 23.20% against *outlier*'s 22.20%, an asymmetry in the direction that distractor-by-editing would produce. We report that ordering with a caveat we would not have attached before auditing it: the similarity score is tied across every option on 32.4% of this corpus's items, typically because the options share no content token at all, so on those items it is the family's index ordering and not the similarity that picks; and under the opposite ordering the same rule reads 18.80% rather than 23.20%. The same audit finds a larger movement still on SHORT (6.00 points), and it reaches the rule that attains this corpus's bound: OVERLAP is tied on 83.0% of items and reads 24.00% under the other ordering against the 27.40% reported here. Appendix G gives the measurement for every rule and corpus, and §5 says what it does and does not change. The ordering between the pair is therefore a weak reading, and the reason to print both is that which one fires is a property of how a corpus's distractors were written rather than a constant, not that either number licenses a conclusion about a corpus on its own.

E LIMITATIONS AND RECOMMENDATIONS, IN FULL

E.1 LIMITATIONS

(0) The bound cannot acquit, and this is the whole shape of the result. Proposition 1 is not a caveat about the implementation. Every number here supports statements of the form "at least this much is blind-solvable", and a small β on some corpus is not evidence that the corpus requires its modality. We have tried to keep the paper's prose inside that shape and would regard any sentence of ours that acquits a corpus as an error.

(1) 8 corpora, not a sample. They were chosen for the regimes they cover and the claims they contest, not because they are representative of anything, and there is no sense in which 8 of them support an inference about benchmarks as a population. They are also concentrated in one family: every row of Table 1 is video or embodied question answering, so a reader is entitled to ask whether what is being measured is a property of blind baselines or a property of that family. The one datapoint we have outside it is the NLVR2 control of Appendix G, which is not a video corpus at all (it pairs two photographs with a written sentence to be judged true or false), and on which the same estimator, the same permutation null and the same refinement sweep run without modification and separate the corpus's two splits by +43.9 points at $g{=}0$. That the instrument transfers to a different task format at all is evidence; one corpus of evidence is not a class-level claim, and a systematic sweep over image-only and text-only benchmark families is the obvious next study rather than something this paper establishes. What has changed since the first version of this section is the kind of statement the breadth licenses: a claim that holds on 5 of 5 option-regime corpora, one of which was built specifically to resist it, is a different object from the same claim on one. The remaining obstacle is unchanged and is not cost: several obvious candidates, SCANQA among them, do not publish annotations separately from gated media, and the endpoint we tried for it during this work serves a re-upload of unclear provenance, which is worse for an auditing instrument than an absence. `analysis/a2_fetch.py` makes adding the next corpus a table entry.

(2) The template family is one choice among many. Granularity is defined by leading question tokens because that ordering is corpus-independent, and Fig. 12 sweeps a trailing-token alternative on the same corpus: the raw estimate falls throughout under it, while the excess still peaks in the interior. Summarised across the two families rather than compared curve to curve, the separation is quantitative. Three summary statistics of the excess (its peak, its mean over the sweep and its

floor) agree between the leading-token and trailing-token families to 0.14, 0.44 and 0.00 points (5.33/3.51/1.56 against 5.47/3.07/1.56), while the raw estimate's mean over the same sweep differs by 2.08 points (6.48 against 4.40) and its shape changes from non-monotone to monotone. That is a two-family robustness check on one corpus and not an invariance result: a template family built from a parser or an embedding would group differently and there is no reason to think the curve's shape is invariant to that, nor that two families agreeing on one corpus makes them agree on another. What Theorem 1 guarantees is the null-induced positive baseline ν_g at every granularity where the corpus's answers repeat, and the collapse at fine granularity. It does not guarantee that $\hat{\beta}_g$ rises between them: that is Proposition 7, which shows the rise is possible, and Table 1, which shows 4 of 9 rows do not take it. We claim neither the peak's location nor its existence as a property of any corpus we have not measured.

(3) The option-regime family is small, and it was widened after the first two corpora. It began at five rules chosen before looking and now holds 9. Doubling it moved the bound by at most 0.33 points on any corpus (0.00 on EGOSCHEMA, 0.17 on VSI-BENCH (choice), 0.33 on PERCEPTIONTEST; §4.1), which is evidence that the family is not badly incomplete and is not proof: by Proposition 1 a family that grew a little tells us nothing about the strategies outside it, and §4.1 reports one strategy outside it that would gain about eight points on the largest corpus. The widening also means the option column is not an out-of-sample reading on the corpora that prompted it. The direction of the incompleteness is known, which is the only property that lets a loose bound be useful.

(3b) The rules' tie-break carries part of what their names claim, and we measure rather than repair it. A surface rule scores each option and takes the argmax; where the top score is tied the implemented family resolves toward the largest option index (§C.4). Ties are not rare. Appendix G re-scores every rule on every option corpus under both orderings: the largest movement anywhere is SHORT on EGOSCHEMA, tied on 37.0% of items and worth 6.00 points between the two (23.00% against 17.00%). Three consequences, and they do not all point the same way. The *bound* is unaffected in validity: each ordering is fixed and blind-visible, so the maximum is attained by a named single-output strategy either way (Proposition 1(i)), and by (ii) enumerating both could only raise it. The *significance* verdict is unaffected in kind: the family-wise null of Table 2 moves each gold to a uniformly random slot, so it carries the same ordering and the same positional advantage, and the comparison is like for like. What is affected is *attribution*, and that part is separable rather than lost. An item's top score is either unique, in which case the score chose, or it is not, in which case the index ordering chose among options the score could not separate; the two buckets' hit counts sum to the rule's own accuracy, so the split is a decomposition and not an estimate. Appendix G reports it for the rule that attains each corpus's bound, and on EGOSCHEMA the larger half of the excess is positional (+5.60 of 27.40%'s margin against +1.80 from the score, §4.1). We therefore do not read the ordering between CENTROID and OUTLIER as evidence about how a corpus's distractors were written, and a reader should not read OVERLAP's margin on EGOSCHEMA as a question–option overlap effect. We did not re-break the ties to make this go away: the five original rules were frozen before any option corpus was read (§G), and amending them after seeing which corpus it helps is the thing the frozen family exists to prevent.

(4) The split estimator's coverage is a real limit, not a footnote. Proposition 4 makes it unbiased on the groups it can score, and at fine granularity those groups are a small, self-selected subpopulation, 3.1% at $g{=}8$. We report coverage in every column and decline to read the estimate where coverage is small; a reader who ignores that column will read the last row as a finding, and it is not one.

(5) The instrument says nothing about models. A visual-dependency gap measured across many models (Lee, 2026) answers a question annotations cannot reach, namely whether a particular system uses the channel it was given. We measure what the corpus permits. The two are complements and the substitution runs only one way: our result bounds theirs from below, and theirs cannot be recovered from ours.

(6) Announcement is a surface test. It counts verbatim containment of content tokens, so it misses paraphrase and counts coincidence. Both errors are present and we have not bounded either; the count is reported as a lower bound on stated answers rather than as a measure of leakage. The stopword list, given in Appendix G, is the one tuning surface in the whole instrument and it was fixed before any corpus was read.

(6b) Two threat models, stated. The blind view of Definition 1 includes the corpus grouping label ℓ alongside the question. An evaluation-time system need not see ℓ: the instrument's estimand is therefore corpus-visible (Threat Model B, pre-release auditing by whoever holds the annotations), not evaluator-visible (Threat Model A, what a deployed blind model sees). The behavioural validation of §4.4 is the bridge between them: the OVERLAP mask on EGOSCHEMA reads no answers and no labels, and the $g{=}1$ modal mask on OPENEQA is a function of the corpus. Reporting which threat model each mask belongs to is what keeps the audit honest.

(7) One positive control is one positive control. §4.2 validates the instrument against a defect that is severe, local, and of a particular kind: a constant gold in a small subset. It does not establish sensitivity to a diffuse defect, to a defect in the option regime, or to one that a template family cannot express. It establishes that the instrument fires when a known defect is present and is silent when the same corpus is repaired, which is more than we could say before and less than a full sensitivity analysis.

(7a) The permutation budget is $R = 1000$, and the convergence data justifies the choice. The instrument uses $R = 1000$ uniform re-deals for the permutation null. The convergence run (`results/a2_nullconv.json`) shows the null mean at $g{=}0$ on OPENEQA moving from 7.73 at $R = 20$ (SE 0.25) to 8.29 (SE 0.03) at $R = 1000$, a shift of 0.56 points (2.2× the $R = 20$ SE). That shift is concentrated at the coarse granularities and is of the same order as the smallest corpus-level excesses reported here (TempCompass: 1.06 points), so the main results use $R = 1000$ throughout: the budget at which the Monte Carlo SE is below 0.05 points at every granularity. The precise claim is about Monte Carlo *error* and not about bias (the R-average is unbiased for the permutation expectation at any R, and a single convergence run shows a realised deviation rather than a displaced estimand), so what $R = 1000$ buys is that the deviation is smaller than anything the paper reads, not the removal of a systematic error we have demonstrated.

(7b) The behavioural validation is a correlation, not an intervention. §4.4 shows that a blind model solves flagged items more often than unflagged ones. It does not show the model is using the flagged regularity: an unobserved third property could make an item both flaggable and easy. The intervention now exists in two halves, and we state exactly where each stands. The annotation-side half is done: the max-statistic permutation null of Table 2 *is* an intervention on the gold's position: every gold is moved to a uniformly random option slot, which destroys the question–gold association and the gold skew together, and the option column's excess is read against that intervention rather than against a hope. The model-side half is now run on OPENEQA, on the 2 strongest blind backbones, under the pre-registered leading-token map of `scripts/a2_interventions.py`: §4.5 and Table 9. It returned a null on both backbones (the gap goes from 37.0–46.1 points to 34.2–49.6, and the stratified differences in differences, -1.0 and +3.6, are indistinguishable from zero, $|z| \leq 1.03$, and carry opposite signs), so what it licenses is a *bound* on the leading token's contribution and not a finding that the contribution is zero. Three things fix the size of that bound, and all three were settled before the result was read. The perturbation's severity is correlated with the mask, so the contrast is computed within severity strata rather than over the corpus (§4.5). The flagged arm is 194 items, so the minimum detectable difference in differences is 11 points pooled: large enough to see a collapse of the 13.7–46.1 point gaps, not large enough to see an erosion of them smaller than about 24–29% of the gap, which is the honest content of the null. And it perturbs one surface feature, so it speaks to the leading token and to nothing else: the option-shuffle half on EGOSCHEMA is implemented in the same script and has not been run. The dose heterogeneity of the first design has been retired rather than argued about: §D.9 reruns the same contrast under a map that replaces every leading token with exactly one token of the same syntactic category, so the word-count difference is zero on every treated item and severity is orthogonal to the mask by construction. That arm also repairs a grammaticality defect the first one had: a leading-token substitution does not undo subject–auxiliary inversion, so the polar questions, which are most of what the mask flags, had been asked ungrammatically. It returns the same null, with both backbones now agreeing in sign and a slightly tighter bound (20–26% against 24–29%). What remains unaddressed is coverage rather than dose: leading tokens with no one-token grammatical equivalent are not treated, so 1432 of 1,636 items carry the intervention. What the existing runs already support without new calls: the ten in/out z-tests are Holm-corrected over the whole family with the corrected separation count reported (Table 5), the negation rates of the two arms are printed beside them so the contrast can be checked against a negation artefact, the OPENEQA mask is re-derived at three granularities and the separation re-tested against the same stored predictions (Appendix G), and the backbones' chosen-index distributions are measured against uni-

form in `results/a2_interventions/baseline_position_bias.json`, which is the pre-registered baseline the shuffle rerun compares against. The behavioural validation still runs on 5 backbones and two corpora (the audit reaches 9 rows but the validation does not, because that one costs interface time), and the EGOSCHEMA result already shows the conclusion depends on which blind strategy the model happens to use.

(8) The main positive control is on our own corpus, and the arm's-length one is weaker. That the parent corpus is ours is what makes its defect independently documented, its mechanism readable off the generator, and its two versions identical outside the repair; it also means the control is not an arm's-length test. Appendix G supplies one that is: NLVR2 (Suhr et al., 2019) publishes an unbalanced split and a balanced one its own authors built to remove a signal exploitable without the images, and the instrument separates them by $+43.9$ points at $g{=}0$ and $+9.9$ at $g{=}1$ while the marginal answer rate moves by $+0.6$. It is weaker than the parent-corpus control in one respect that matters: the defect there is a documented design property rather than a hand-found bug with a known line of generator source, so what is being calibrated against is the authors' intent rather than a mechanism we can read. Both are reported and neither is a sensitivity analysis.

E.2 RECOMMENDATIONS

Three, addressed to anyone reporting a blind baseline, and all of them cheap.

(i) State the granularity and the null. A blind-baseline number without the grouping it was computed at and the null it was read against is a number without a referent. Theorem 1 makes the null-induced baseline and the collapse unavoidable, and Proposition 7 makes the intermediate rise available, so both directions of error are reachable by changing a choice that is usually left implicit; and which direction a given corpus takes cannot be predicted, only measured (5 of 9 rows here are non-monotone, 4 are not). Report the null alongside the estimate, and report the estimate too: §4.2 shows a defect the excess cannot see and β can.

(ii) Score with a split, not leave-one-out, and report coverage, by items as well as by groups. The split estimator moved 0.84 points across granularities where leave-one-out on the same groups moved 1.91 and leave-one-out over the whole corpus moved 5.14 (§4.3). It costs nothing extra and it removes most, not all, of the refinement trap; the part it removes by restricting the population rather than by scoring better is why the coverage columns have to be printed.

(iii) Do not attach a bootstrap interval. Proposition 3 and §4.3 show it reinstates the leak the held-out estimator exists to avoid, inflating β by up to $6.96\times$. Report the permutation null, which is exact for the observed answer multiset and resamples no items.

E.3 CONCLUSION

The finding that a multimodal benchmark does not need its modality has been made repeatedly, each time by running systems on a published corpus. A substantial part of that finding is a property of the corpus and is computable from its annotations, on a CPU, with no model and no frames, which is what makes it runnable on corpora whose pixels are behind an agreement. The instrument condemns and never acquits, and we prove why no instrument of this kind can do otherwise.

We validate it where an instrument paper usually cannot: against a defect that was found by hand and documented before the instrument existed, on two corpus versions that are identical outside the repair. It reports 100.00% before and 0.00% after, and a corpus-wide excess over the permutation null of 12.13 against 3.14 points. A second control runs the same estimator across a repair its own authors performed (Appendix G).

The result we did not expect is about the estimator rather than about the corpora. Blind solvability estimated inside groups of similar questions carries a positive baseline set by the group-size profile alone, and refining the grouping destroys the estimate, on 5 of 9 rows here after moving it up first, though that rise is a possibility we construct and measure rather than a consequence we prove. Behind it sit two leaks and an overfit: the answer being scored returns to its own evidence in the in-sample estimator and in the bootstrap, and stands in for itself through a single neighbour in leave-one-out on a group of two. The bootstrap is the dangerous one, because it is what a careful author reaches for and it inflates the estimate by up to $6.96\times$ while looking like rigour. We diagnose it by deduplicating

its own draws, and we supply the repair: a fit-and-test split, which is flatter than leave-one-out on the same groups and flat in absolute terms where it has coverage.

Put together, the three results are one object rather than three findings about benchmarks. The first is a *measurement*: a blind-exploitability lower bound computable from annotations alone, attained by a strategy we can name and print. The second is a piece of *measurement theory*: that such a bound is one-sided by necessity rather than by implementation, so it can condemn a corpus and can never clear one, and that no finite family of strategies closes the gap. The third is a catalogue of *measurement pathology*: the granularity a bound is computed at, the leak in its estimator, the overfit beside the leak, the bootstrap interval that reinstates the leak, the tie-break hidden inside a rule's name, and the selection hidden inside a maximum over rules; each of these moves a published number by more than the effects such numbers are normally used to argue about. An audit protocol is what the three make when they are used together, and the corpora here are where we checked that it works rather than what it is for.

Whatever else is taken from this paper, a reported blind baseline should carry the granularity at which it was computed, the null it was read against, the family its maximum was taken over, and the coverage of its estimator, or it is a number without a referent, and possibly a number about its own error bars.

F Full Proofs

Standing conventions

A corpus is a finite multiset of items $x = (q, a, \ell)$ as in Definition 1, with N items and answers drawn from a finite set $\mathcal{Y}$. We write p for the corpus-wide empirical answer distribution, τ_g for the template at granularity g, $\{n_i\}_{i=1}^{G(g)}$ for the induced group-size profile, and $\hat{y}(M)$ for the deterministic single-valued mode of a multiset M: the most frequent element, ties broken by a fixed blind-visible ordering (lexicographically smallest normalised answer string; smallest gold index in the option regime; $\hat{y}(\varnothing)$ undefined). The leave-one-out modal estimator scores a hit on item x in group P when $\hat{y}\big(\{a_y : y \in P, y \neq x\}\big) = a_x$, which is false when the remainder is empty. All permutation statements are with respect to a uniformly random bijection of the corpus-wide answer multiset onto the item set, holding the profile fixed. The estimator is therefore exactly the accuracy of one named single-output blind strategy, and Proposition 1(i) holds as stated.

Proposition 1 (one-sidedness)

(i) Immediate: $\hat{\beta}$ is a maximum over a finite set, so it is the accuracy of some $\sigma \in S$, and every member of S is written out in Definition 10 or §3. The free-form modal member is the single-valued rule with the fixed blind-visible tie-break stated in §3 and the appendix conventions, so the maximiser is one named single-output blind strategy per corpus, not an oracle choice among tied modes. Nothing is inferred; a reader can recompute the attaining rule on a single item by hand.

(ii) Each $\sigma \in S$ belongs to $\mathcal{C}$, so its accuracy is one of the values over which $\beta^\star(\mathcal{C})$ takes a supremum, whence $\hat{\beta} = \max_{\sigma \in S} \operatorname{acc}(\sigma) \leq \sup_{\sigma \in \mathcal{C}} \operatorname{acc}(\sigma) = \beta^\star(\mathcal{C})$. Note that this holds however the maximising member of S was selected: the selection is over accuracies of members of the class, not over the class itself.

(iii) We exhibit the corpus. Write $S = S_0 \cup S_1$, where S_0 are the rules that read no answers at all (fixed maps $\sigma : (q, \ell) \mapsto \mathcal{Y}$, which is what every surface rule of Definition 10 is), and S_1 are the rules fitted inside template groups, of which the modal-answer rule and INDEX are the instances used here. Let $m = |S_0|$ and suppose $|\mathcal{Y}| > m$. Build a corpus of N items whose blind views $v_1, \ldots, v_N$ are pairwise distinct. For each i choose $f(v_i) \in \mathcal{Y} \setminus \{\sigma(v_i) : \sigma \in S_0\}$, which is possible because that excluded set has at most $m < |\mathcal{Y}|$ elements, and set $a_i = f(v_i)$. Then every $\sigma \in S_0$ errs on every item by construction; and every $\sigma \in S_1$ errs on every item too, because each blind view is its own group and the held-out remainder of a singleton group is empty, so no mode exists. Hence $\hat{\beta} = 0$. But f itself is a map from blind views to $\mathcal{Y}$ that reads no answers, so it is a blind strategy by Definition 1, and it is correct on every item. A small $\hat{\beta}$ is therefore consistent with a corpus that is entirely blind-solvable, and the instrument condemns without acquitting. ■

Remark. The construction is not a pathology introduced to win an argument; it is the formal shape of the thing §4.4 measures. There, a blind backbone reaches 47.4% on EGOSCHEMA while the best rule in S reaches 27.4%, because the backbone is using world knowledge: a map from blind views to answers that S does not contain. The gap is the difference between a bound and the quantity it bounds, and Remark 1 explains why that quantity cannot be defined as a supremum over everything.

PROPOSITION 2 (ANNOUNCEMENT IS A DISTINCT DETECTOR)

Let x be an announced item, so every content token of a_x occurs among the tokens of q_x. Then $\mathrm{content}(a_x)$ is one of the sub-multisets of $\mathrm{content}(q_x)$, a family of at most $2^{|q_x|}$ candidates that a reader of the blind view can enumerate without the modality and without any answer. That is the sense in which announcement is a blind-exposure statement: on those items nothing outside the question is needed to *contain* the answer.

What does not follow is an inequality between V and $\hat{\beta}$, and it is worth being explicit about the two independent reasons.

First, containing the gold in an enumerable candidate set is not selecting it. A strategy that returned, say, the first k content tokens of q would be blind, but it would not in general return the gold: on *"what colour is the brown chair?"* with gold `brown` it returns `colour`. Selecting the right sub-multiset is exactly what requires the answer, which Definition 1 forbids. No rule in the paper's S attempts it, so $V \leq \hat{\beta}$ is not implemented and is not claimed.

Second, and decisively, the inequality is false as printed. Table 1 reports $V = 100.00\%$ and $\beta_{g=0} = 24.94\%$ on VSI-BENCH (choice), because the stem of every item enumerates the whole option set and therefore announces the gold and every distractor at once. Announcing all K options tells a blind reader nothing, which is why Definition 12 reports V^{disc} there, and V^{disc} on that corpus is 0.00%.

What the proposition therefore supports is the use the paper makes of it and no more: the two counts detect one kind of failure by two routes, so they must not be added, and the relation between the two rates is a measurement. Measured on OPENEQA, β exceeds V in 6 of the 7 categories and falls below it in 1 (OBJECT LOCALIZATION, $V = 0.76\%$ against $\beta = 0.00\%$; Fig. 8), while the two flag nearly disjoint *items* (§4.6). The exception is worth keeping rather than rounding away: it is the proposition's refusal to order the rates, exhibited inside a single corpus and not only across the two regimes. ■

THEOREM 1 (NULL-INDUCED BASELINE AND COLLAPSE)

(i) If every answer in the corpus is distinct then no remainder ever contains the held-out answer, no deal produces a hit, and $\nu_g = 0$ for every profile; this is why a repeated value is required. Suppose then that some value y occurs at least twice, and take first a group with $n_i = 2$. Under a uniform permutation the event that both copies of y are dealt into that group has positive probability, since every assignment of the multiset to positions is equally likely. On that event the remainder of either item is the single answer y, which is its own deterministic mode, so both items score hits and $\nu_g > 0$.

The same argument works for larger groups whenever some positive-probability deal gives the group a strict mode w of multiplicity $c \geq 2$: removing one copy of w leaves it at $c - 1$, which is strictly greater than the count of every other value, so w is the deterministic mode of the remainder regardless of the tie-break and that item scores a hit. Strictness is needed: on a group dealt $\{a, a, b, b\}$ every item's removal leaves a tied remainder whose deterministic prediction is fixed by the tie-break, so at most one of the two answer values scores, which is also why the post-repair control of §4.2 reads 0.00% rather than chance.

Note what (i) does and does not assert. We *define* the excess as $\hat{\beta}_g - \nu_g$. That it is the part of the estimate attributable to association between template and answer is a modelling choice, not a further theorem; it is made because ν_g is by (ii) exactly the value the same estimator returns when that association is destroyed and nothing else is.

(ii) Under a uniform permutation of the corpus-wide answer multiset, the joint distribution of answers within groups depends only on the group sizes and the multiset, by exchangeability: the permutation is

invariant to any relabelling of items within a group. The estimator is a function of that joint distribution alone (it reads only the multiset of answers in each group), so its expectation is determined by the profile and the corpus-wide multiset, and by no association between template and answer. This is what makes ν_g computable exactly by simulation without a parametric model of p, which matters because answer distributions here are long-tailed (1001 distinct answers over 1,636 items).

(iii) Refining τ_g to τ_{g+1} splits groups and never merges them: two items with distinct keys at g have distinct keys at $g+1$, since the key at g is a prefix of the key at $g+1$. Hence each n_i is non-increasing in g. When all $n_i = 1$ the held-out remainder of every group is empty, no deterministic mode exists, and both $\hat{\beta}_g$ and ν_g are zero. If some $g = 0$ group has a strict mode w of multiplicity $c \geq 2$, then holding out one of its c copies leaves w at $c - 1$, which strictly exceeds every other count, so w is the deterministic mode of the remainder regardless of the tie-break and that item scores a hit; hence $\hat{\beta}_0 > 0$. The hypothesis is on a *group* and it is not implied by the corpus's answers repeating: two groups each holding $\{a, b\}$ repeat both values and, under the fixed tie-break, score zero on every item (the held-out remainder is always the single other answer, which cannot equal the held-out gold), so $\hat{\beta}_0 = 0$ even though both answers repeat; and a single group holding $\{a, a, b, b\}$ scores at most the items on one side of the tie. A sequence that is positive at $g = 0$ and zero at large g cannot be non-decreasing, which is the whole of the monotonicity content available here. It is *not* thereby non-monotone: it could decrease throughout, and on 4 of the 9 rows of Table 1 it does exactly that or the opposite, without turning. ■

PROPOSITION 7 (THE CURVE CAN RISE BEFORE IT COLLAPSES)

It suffices to exhibit one corpus. Take eight items sharing one corpus label ℓ, with answers a, a, a, a, a, b, b, c and a, b, c pairwise distinct with a the smallest in the fixed tie-break ordering. Let the five items answered a have questions beginning with the token `w`, the two answered b have questions beginning with `v`, and the one answered c have its question beginning with `u`, and let all eight second tokens be distinct. This is an admissible template family by Definition 2: τ_0 is ℓ, and each increment adds one leading question token.

At $g = 0$ the partition is the single group $\{a^5, b^2, c\}$. Holding out an a leaves $\{a^4, b^2, c\}$, whose unique mode is a, so each of the five a items scores a hit; holding out a b leaves $\{a^5, b, c\}$, whose unique mode is a, so neither b item does; holding out c leaves $\{a^5, b^2\}$, whose unique mode is a, so c does not either. Hence $\hat{\beta}_0 = 5/8 = 62.5\%$.

At $g = 1$ the partition is $\{a^5\}$, $\{b^2\}$ and $\{c\}$. Holding out any item of the first two groups leaves a remainder whose unique mode is that item's own answer, so seven items score hits; the singleton $\{c\}$ has an empty remainder and scores 0. Hence $\hat{\beta}_1 = 7/8 = 87.5\%$.

At $g = 2$ every group is a singleton, every remainder is empty, and $\hat{\beta}_2 = 0$. So $62.5 < 87.5 > 0$, and the sequence is not monotone. ■

Remark. Lemma 1 supplies the mechanism that makes this shape common rather than contrived: the null falls with group size while the fitted part rises, so their difference peaks in the interior. It does not make the shape universal, which is why this is a proposition about existence and Table 1 is where the question of which corpora take it is settled.

PROPOSITION 3 (TWO LEAKS AND AN OVERFIT)

(1) In-sample (a leak). The multiset from which the deterministic mode is computed contains a_x. On a group of size 1 the mode is a_x itself and the item scores a hit with certainty. Since refinement drives every group toward size 1 (Theorem 1(iii)), the in-sample rate tends to 1 regardless of the corpus.

(2) Leave-one-out on a group of size 2 (an overfit, not a leak). The remainder is a single item y, so the deterministic mode is a_y and the single-valued estimator scores a hit exactly when $a_x = a_y$. The held-out answer is genuinely absent from the evidence, so Definition 3 is not violated; what is wrong is different. The expectation over such groups is the probability that two items sharing a template share an answer, which is a measure of the group's homogeneity, and the "strategy" being scored is a mode fitted to a sample of size one. It is not the accuracy of anything applicable to an unseen item. The cure is the same as for a leak (fit on one part of the group and score on another, Definition 4),

which is why the three are presented together. Under the single-valued rule this case is tie-free, so it is also the cleanest witness that the small-group pathology survives the tie-break fix.

(3) Bootstrap. Draw n items from n with replacement, and let K_x be the multiplicity of a given item. Then K_x is binomial with n trials and success probability $1/n$, which converges in distribution to a Poisson variable of mean 1. Hence

$$\Pr[K_x \geq 2 \mid K_x \geq 1] = \frac{1 - \Pr[K_x{=}0] - \Pr[K_x{=}1]}{1 - \Pr[K_x{=}0]} \xrightarrow[n\to\infty]{} \frac{1 - e^{-1} - e^{-1}}{1 - e^{-1}} = \frac{1 - 2e^{-1}}{1 - e^{-1}} \approx 0.418,$$

which is Eq. (1). Every copy beyond the first lies in the same template group as the held-out copy (the template is a function of the blind view, which is identical across copies), and votes for a_x. On a group whose original size is small, one such vote is decisive, so the leave-one-out estimator computed on a bootstrap draw behaves like the in-sample estimator on the duplicated items. The effect therefore grows as granularity refines, which is what Table 8 measures. ■

Remark on the diagnosis. The argument above is a mechanism, not a measurement, and mechanisms can be wrong. §4.3 tests it by scoring each draw twice: once as drawn, and once with duplicates removed from that same draw and nothing else changed. If the inflation were a property of resampling (of the altered group-size profile, say), removing duplicates would not remove it. It removes it at every granularity, which is what identifies the cause.

LEMMA 1 (ASYMPTOTIC LOCATION OF THE PERMUTATION NULL)

Under the null, the answers in a group of size n are n exchangeable draws taken *without replacement* from the corpus's answer multiset. The statements below are proved for the i.i.d. approximation announced in the lemma, $X_i \overset{\text{iid}}{\sim} p$, which is exact in the limit of a large corpus and is the regime the simulation operates in; for a finite corpus the without-replacement null differs by the usual $O(1/N)$ sampling-fraction correction, which is why the lemma is used to locate the simulated null asymptotically and never to replace it.

$\nu(1) = 0$: the remainder is empty, $\hat{y}$ is undefined, and no hit is possible.

$\nu(2) = \sum_y p_y^2$: the remainder is a single draw X_1, whose deterministic mode is X_1, so a hit occurs exactly when the held-out answer equals X_1; both are independent draws from p, and the collision probability is $\sum_y p_y^2$.

$\nu(n) \to \max_y p_y$: let $y^\star$ be the unique maximiser of p (if the maximiser is not unique the limit is the total mass on maximisers under the fixed tie-break ordering, and the argument is unchanged). By the strong law the empirical frequency of each y among $n-1$ draws converges to p_y, so $\Pr[\hat{y} = y^\star] \to 1$ and $\Pr[\hat{y} = y] \to 0$ for $y \neq y^\star$. Hence $\nu(n) = \sum_y p_y \Pr[\hat{y} = y] \to p_{y^\star} = \max_y p_y$.

Finally $\sum_y p_y^2 \leq (\max_y p_y) \sum_y p_y = \max_y p_y$, so $\nu(2) \leq \nu(\infty)$ and the null is non-decreasing across the three stated points. ■

PROPOSITION 4 (THE SPLIT ESTIMATOR DOES NOT LEAK)

Fix a group P with $|P| = n \geq 2$ and a partition into F (fit) and T (test) with $F \cap T = \varnothing$ and $F \neq \varnothing$. The prediction for every $x \in T$ is the deterministic single-valued mode $\hat{y}(\{a_y : y \in F\})$ with the fixed blind-visible tie-break, a function of F alone, and $x \notin F$; hence a_x does not enter the multiset from which its own prediction is formed, which is the negation of Definition 3. This holds for every $n \geq 2$, with no dependence on group size, which is the property leave-one-out lacks.

For what it estimates, the honest statement is narrower than the one this proof used to make. Define θ_g as the expected item-weighted accuracy, over uniformly random partitions of every group with $n_i \geq 2$ into a fit half of $\lfloor n_i/2 \rfloor$ and a test half of the rest, of the modal rule fitted on the fit half and applied to the test half. The estimator of Definition 4 is an average of exactly this quantity over independent random partitions, so it is unbiased for θ_g by construction. It is unbiased for nothing else, and in particular θ_g is not the accuracy of a mode fitted on $n_i - 1$ items, which is what leave-one-out targets: a mode fitted on half a group is fitted on less evidence and is a different rule. Nor is θ_g the accuracy a blind system would attain on *fresh* items of the same template, because there are no such

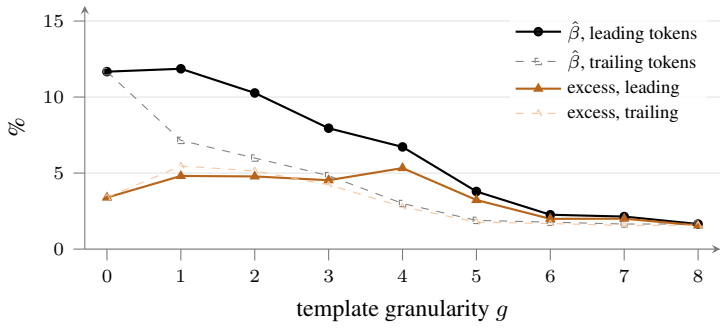

Figure 12: Two template families on the same corpus. Cutting questions by leading tokens gives a non-monotone $\hat{\beta}$ that rises before collapsing; cutting by trailing tokens gives one that falls throughout, from 11.7% to 1.6%. *The excess peaks in the interior in both*, at 5.33 points (g=4) and 5.47 points (g=1). Summarised over the whole sweep rather than at the peak, the three statistics of the excess agree across the two families to 0.14 (peak), 0.44 (mean over g) and 0.00 (floor) points, while the raw estimate's mean differs by 2.08 points and its *shape* changes outright. So the shape of the raw estimate is a property of the template family and the excess is substantially not, which is the concrete reason §5 declines to claim an invariant peak location and §D.3 reports the excess.

items: the group is all there is, and an exchangeability argument over items that do not exist proves nothing. What can be said is that the test items played no part in fitting the rule that scores them, which is the property leave-one-out lacks at small n_i and the only property the repair claims.

For coverage: a group with $n = 1$ admits no partition with both parts non-empty and is excluded, so the covered set is $\{i : n_i \geq 2\}$ and the coverage is $|\{i : n_i \geq 2\}|/G(g)$. By Theorem 1(iii) group sizes are non-increasing in g, so coverage is non-increasing in g and reaches 0 when every group is a singleton. The estimator is unbiased on the covered subpopulation and says nothing about the excluded one, and the excluded one is selected precisely by having an uncommon template; the covered subpopulation is therefore not a uniform sample of the corpus, and the estimate must be read with its coverage. ■

G Secondary Experiments and Sensitivity

This appendix holds five things the main text refers to and does not have room for: the alternative template family (Fig. 12), a worked group showing how a hit is scored (Fig. 13), the division of labour between this instrument and the model-side diagnostics it complements (Fig. 14), the full option-regime family (Table 12), and the third-party positive control on NLVR2. It closes with a note on what was fixed before the data were seen and what was not (§G).

The full blind-strategy family

§4.1 reports that widening the option-regime family from five rules to 9 moved the bound by 0.00 points on EgoSchema and by up to 0.33 points elsewhere. The whole family is given here, because a bound that is a maximum over an enumerated set is only as auditable as the set. Fig. 11 plots the original five and Table 12 lists all 9.

Is the mask an artefact of the granularity it is read at?

The OpenEQA mask of §4.4 is the leave-one-out modal hit set at g=1, and g=1 is exactly where §4.3 says the small-group overfit begins. Validating an instrument with a mask drawn from the regime the same paper calls unstable is a fair objection, and it has a cheap answer, because the objection is about the *partition* and not about the model runs: the stored blind predictions are per item, so re-deriving the mask at another granularity and re-splitting the same predictions costs no call and no GPU. Table 13 does that at g=0, g=1 and g=2.

The three granularities are genuinely different partitions rather than three names for one. At g=0 the mask is built from the corpus's own category label alone, over 7 groups with a median of 231 items

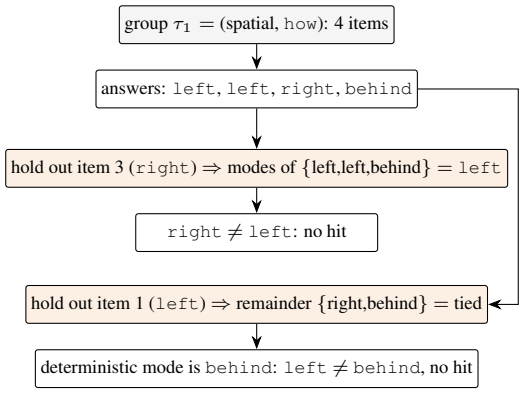

Figure 13: One group, worked, under the fixed blind-visible tie-break of §3. Holding the item out is what makes the number an estimate rather than a description: an in-sample modal predictor scores near one on any corpus whose answers are distinct. The second case shows the tie-break: when the remainder is tied, the deterministic prediction is the lexicographically smallest answer (`behind`), which the held-out item (`left`) does not match, so it scores no hit. This is what makes the bound a single-output strategy rather than an oracle choice among tied modes.

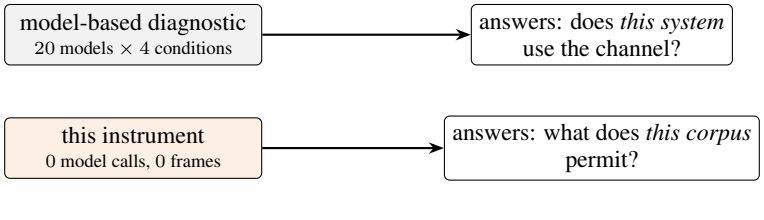

Figure 14: The two kinds of evidence and what each can answer. A visual-dependency gap measured across many models (Lee, 2026) reaches a question about systems that annotations cannot, and costs a full evaluation grid. The instrument here costs nothing and answers a question about the corpus. The paper claims the second, not a replacement for the first.

and 0.0% singletons; at g=2 it is built over 280 groups with a median of 1 and 60.0% singletons. The flagged arm moves accordingly, from 191 items at g=0 to 194 at g=1 and 168 at g=2, because the leave-one-out modal rule abstains on a singleton and singletons multiply as groups shrink.

The separation does not move with it. All 15 of the 15 cells put the flagged arm above the unflagged one, 15 survive a Holm step-down over the whole sweep, and the lift runs 1.56–$5.44\times$ against 2.31–$4.64\times$ at the reported granularity. What this rules out is the specific failure the objection names: that the behavioural separation is manufactured by the small groups g=1 produces rather than by a property of the corpus. What it does not do is turn the mask into a granularity-free object: *which* items are flagged changes with g, and that is Theorem 1 rather than a defect. The claim the sweep supports is the one §4.4 makes: the flagged set is not an artefact of its estimator, at any of the three granularities where the estimator has enough group left to speak.

What the tie-break is worth

§C.4 states that the surface rules resolve a tied top score toward the largest option index and that QLENMATCH and INDEX resolve toward the smallest. Table 14 is what that is worth. It re-scores every rule on every option corpus under both orderings, changing nothing else; the reimplementation is checked against the instrument's own numbers under the ordering the instrument uses, and a disagreement is fatal rather than reported.

Table 15 is the sharper statement and the one §4.1 quotes. Rather than asking how far a number would move under the other ordering, it asks how much of the number the score was responsible for

Table 12: The full option-regime family, reported including the rules that fail. Enumerating in advance and reporting all of them is what keeps $\hat{\beta}$ a bound rather than a selection effect: a family searched until something exceeded chance would guarantee an apparent artefact on any corpus. Four rules were added after the original five and the bound did not move *on this corpus* (§4.1), which is evidence that the family is not badly incomplete and is not proof, since Proposition 1 is not repealed by a family that failed to grow. Two losers are informative: a constant index beats chance, and CENTROID beats OUTLIER (23.20% against 22.20%), the asymmetry that writing distractors by editing a correct answer would produce. On two of the other option-regime corpora one of the four added rules does win, by 0.17 and 0.33 points, and Table 1 reports the widened family throughout.

Rule	what it picks	accuracy (%)	vs. chance
OVERLAP	most content-token overlap with the question	**27.40**	+7.40
INDEX	modal gold index within a template group (leave-one-out)	23.40	+3.40
FIXED	a single constant index, the best one	23.40	+3.40
CENTROID	most similar to the other options	23.20	+3.20
SHORT	the shortest option	23.00	+3.00
OUTLIER	least similar to the other options	22.20	+2.20
QLENMATCH	length closest to the median option length	14.20	-5.80
LONG	the longest option	13.60	-6.40
UNIQ	most tokens absent from the other options	12.60	-7.40
chance, $1/K$ with $K = 5$		20.00	—

Table 13: Mask-granularity robustness on OPENEQA. The mask of Table 5 is the leave-one-out modal hit set at g=1, which is also the granularity at which §4.3 says the small-group overfit begins, so this table re-derives the mask at g=0, g=1 and g=2 and re-splits the *same* stored blind predictions by it. No model is re-run and no call is made: only the partition of the items changes, which is what makes this check cost nothing. n_{in} is how many items the mask flags at that granularity; it falls as groups shrink, because the leave-one-out modal rule abstains on singletons. IN and OUT are accuracy inside and outside the mask, within a corpus and within a model. p_{Holm} steps down across the whole 15-cell sweep (3 granularities $\times$ 5 backbones), so the robustness check is corrected on the same footing as the headline table rather than quoting 15 raw p-values. What is being checked is the direction and the survival, not the point values: 15 of 15 cells separate the arms in the same direction and 15 survive the step-down, so the behavioural separation of §4.4 is not an artefact of the one granularity the reported mask is read at.

g	Backbone	n_{in}	IN (%)	n_{out}	OUT (%)	lift	p_{Holm}
0	Ministral 3B	191	**51.3**	1433	11.6	4.43	<0.001
	Qwen3-VL-8B	191	**47.6**	1433	14.0	3.41	<0.001
	Gemma 3 12B	191	**17.8**	1433	11.4	1.56	0.011
	Qwen3.8-27B	191	**64.9**	1433	11.9	5.44	<0.001
	Gemma 4 31B	191	**29.8**	1433	9.4	3.17	<0.001
1	Ministral 3B	194	**51.5**	1430	11.5	4.49	<0.001
	Qwen3-VL-8B	194	**50.5**	1430	13.5	3.74	<0.001
	Gemma 3 12B	194	**24.2**	1430	10.5	2.31	<0.001
	Qwen3.8-27B	194	**58.8**	1430	12.7	4.64	<0.001
	Gemma 4 31B	194	**30.4**	1430	9.3	3.27	<0.001
2	Ministral 3B	168	**49.4**	1456	12.4	3.97	<0.001
	Qwen3-VL-8B	168	**48.8**	1456	14.4	3.40	<0.001
	Gemma 3 12B	168	**23.2**	1456	10.9	2.14	<0.001
	Qwen3.8-27B	168	**61.3**	1456	13.2	4.65	<0.001
	Gemma 4 31B	168	**25.6**	1456	10.2	2.50	<0.001

at all, which is answerable exactly, because an item's top score either separates the options or it does not, and the two buckets' hit counts sum to the rule's accuracy.

THE CONTROL'S MECHANISM, READ OFF THE GENERATOR

§4.2 rests on the claim that the two corpus versions differ in one thing only. The claim is checkable rather than asserted, because the corpus is procedurally generated from a seed and the generator is available (Anonymous authors, 2026c).

Table 14: How much of each option-regime rule is decided by its tie-break rather than by its score, on every option corpus. A surface rule scores each option and takes the argmax; where the top score is tied, the implemented family resolves toward the *largest* option index and QLENMATCH and INDEX toward the smallest. Both are fixed and blind-visible, so every rule is a single-output blind strategy either way and Proposition 1(i) is untouched; the column that matters is the last one, which is how far the reported accuracy would move under the opposite convention. TIED is the share of items whose top score is not unique: on this family a top tie always spans at least two indices, so it is also the share the convention decides. Rules are ordered by swing and the table is truncated to the rows above 1.00 points; the full grid of 35 cells is in `results/a2_tiebreak.json`. Read with Table 2: the family-wise null moves each gold to a uniformly random slot, so it carries the same tie-break and the same positional advantage, and the significance verdicts are therefore comparisons at a fixed convention.

Corpus	rule	tied (%)	as reported	other convention	swing
EGOSCHEMA	SHORT	37.0	**23.00**	17.00	6.00
	CENTROID	32.4	**23.20**	18.80	4.40
	OVERLAP	83.0	**27.40**	24.00	3.40
	QLENMATCH	48.6	**14.20**	16.60	2.40
VIDEO-MME	OVERLAP	86.1	**22.56**	24.37	1.81
TEMPCOMPASS	QLENMATCH	93.7	**31.79**	30.02	1.77
VIDEO-MME	QLENMATCH	100.0	**24.00**	22.59	1.41
TEMPCOMPASS	UNIQ	49.6	**29.45**	28.06	1.39
VIDEO-MME	LONG	52.8	**28.15**	29.48	1.33
	CENTROID	75.0	**26.19**	27.52	1.33
TEMPCOMPASS	OVERLAP	84.8	**30.21**	29.20	1.01
VIDEO-MME	UNIQ	54.1	**25.30**	26.30	1.00

Table 15: The rule that attains each corpus's option-regime bound, split into the part its score decided and the part its tie-break decided. An item's top score is either unique (and then the score chose) or it is not, and then the index ordering chose among options the score could not separate; the two buckets' hit counts sum to the rule's own accuracy, so this is a decomposition and not an estimate. SCORE and TIE give each bucket's share of the corpus and its contribution to the rule's excess over chance, and the two contributions sum to that excess. The reading is not that the bound is wrong: both halves are blind, both are inside the enumerated family, and neither needs the modality, so Proposition 1(i) is untouched. The reading is that a rule's name is not evidence about which regularity a corpus carries, and on the corpus where the split is most lopsided the larger half is positional rather than lexical.

Corpus	winning rule	β	chance	SCORE (% / exc.)	TIE (% / exc.)	all-tied
EGOSCHEMA	OVERLAP	27.40	20.00	17.0 / +1.80	83.0 / +5.60	60.8%
VSI-BENCH (choice)	QLENMATCH	28.88	25.00	0.0 / +0.00	100.0 / +3.87	89.8%
VIDEO-MME	LONG	28.15	25.00	47.2 / +3.83	52.8 / -0.69	34.3%
TEMPCOMPASS	LONG	34.39	33.33	43.3 / +0.95	56.7 / +0.10	29.2%
PERCEPTIONTEST	QLENMATCH	36.80	33.33	29.6 / +2.23	70.4 / +1.23	35.9%

The order-comparison item asks "did you see the X before or after the Y?" for two objects each appearing exactly once. In the earlier version the generator drew the two objects from a list returned in sorted frame order and named them in that order, so the earlier object was always named first and the gold was always BEFORE. The repair does not randomise the draw (that would move the shared random stream and change every downstream sample, regenerating the whole corpus), but derives a fixed four-BEFORE/four-AFTER assignment from the corpus seed and consumes it by slot index, then chooses the naming order to realise the assigned gold. Because the assignment is drawn from an item-local generator, the frames, the singleton objects, the tracked classes and the chain items are bit-identical between the two versions. We verified that by comparing the shipped metadata field by field.

Two properties of this repair matter for reading Table 3. The balance is *constructed* rather than sampled: eight independent coin flips would land on four-four less than a third of the time, and a six-two split still lets a constant answer score 75% on a subtest whose chance level is meant to be 50%. And the defect it removes is the strongest kind a blind bound can detect (a constant gold), which is why the control demonstrates sensitivity to a severe local defect and not sensitivity in general (§5, item 7).

THE ANSWER DISTRIBUTION, AND WHY THE NULL IS WHERE IT IS

Lemma 1 bounds the permutation null by two quantities computed from the corpus-wide answer distribution p alone. On OPENEQA there are 1001 distinct normalised answers over 1,636 items, with collision probability $\sum_y p_y^2 = 1.54\%$ and base rate $\max_y p_y = 8.80\%$.

The measured null runs from 8.29% at $g{=}0$ to 0.09% at $g{=}8$. The first sits just below the base rate, which is what Eq. (2) predicts for large groups. The second sits *below* the collision probability, which is also predicted, and for a reason worth spelling out: the corpus-level null is an average over groups, and at $g{=}8$ almost every group is a singleton with $\nu(1) = 0$, so the average is pulled toward zero rather than toward $\nu(2)$. Both ends therefore land where the lemma says, and neither was fitted.

THE DUPLICATE RATE, SIMULATED

Equation (1) is asymptotic. At the corpus size used here the simulated share of distinct items appearing at least twice in a bootstrap draw is 0.418 against the asymptotic 0.418, so the finite-sample correction is negligible and the mechanism of Proposition 3(3) applies at the sizes that occur in practice. The inflation it produces is not negligible: it grows from $0.99\times$ at $g{=}0$, where groups are large enough that one duplicate vote rarely carries the mode, to $6.96\times$ at $g{=}4$, where most groups are small enough that it always does.

WHAT WE WOULD RUN NEXT, AND WHAT STOPPED US

The instrument's cost is negligible (Proposition 11) and its breadth is limited by annotation availability rather than by compute.

Two intake failures are on the record, and §D.1 refers to them. In the first, an attempt to add a third open-answer corpus retrieved nothing from either annotation endpoint tried, and the files left behind were fourteen bytes each: the JSON body of a 404 page. Fourteen bytes of JSON parses, so the corpus was simply absent from a study that recorded having tried. Nothing downstream complained, and nothing would have. The replacement fetcher checks every download against the serving endpoint's own row count and refuses a short file, which is why the second failure was caught: PERCEPTIONTEST arrived with its `answer_id` field serialised as the string `"2"` rather than the integer 2, an `isinstance` guard dropped all 19140 rows, and the empty corpus surfaced as a division by zero in the announcement count rather than as a silent absence. A pipeline that fails loudly on an empty input is the only reason either of these was found.

What remains is the corpus the earlier attempt was aimed at. SCANQA does not publish its annotation files separately from gated media; the mirrors that exist are third-party re-uploads of unclear provenance, and for an instrument whose entire claim is that it audits a corpus's own annotations, auditing somebody's copy of them is worse than not auditing.

The gap SCANQA was meant to fill has since been filled from elsewhere, and we record what that changed, because it is the one place where the paper's own stated next step was taken inside the paper rather than left to a reader. Five of what were then seven rows were option regimes, so the free-answer estimator rested on fewer corpora than the option-regime bound did. Two open-answer corpora were added (ACTIVITYNET-QA (Yu et al., 2019) and the open-ended split of NEXT-QA, Xiao et al., 2021), taking Table 1 to 9 rows and 47863 items, of which 4 rows are free-answer rather than two.

Both exceed their own permutation nulls, and one does so by more than any free-answer corpus previously in the table: the largest free-answer excess is now 13.23 points against 5.33 before. That is worth stating precisely because both corpora are open in *format* and substantially binary in *content* (a large share of their answers are `yes` or `no`), so the blind bound registers a high base rate. That is a property of the corpora and is reported rather than corrected for; it is also the reason the announcement count V stays near zero on both while β does not, which is the ordering Proposition 2 predicts and the one an announcement-only audit would miss.

Two studies this paper does not run, and why each is a next paper rather than a missing section. We name them here so that the boundary of the present result is explicit rather than inferred. A third,

the matched-dose leading-token intervention, was named here in an earlier version and has since been run: it is §D.9.

A cross-family sweep. Every row of Table 1 is video or embodied question answering; NLVR2 is the one datapoint outside that family and it is a control rather than a survey. Whether the refinement curve's shape, the null-induced baseline's size and the option-regime family's coverage behave the same way on image-only and text-only benchmark families is a question this paper's theory applies to and its measurements do not answer. It is a corpus-intake problem rather than a compute one (`analysis/a2_fetch.py` makes each addition a table entry), and it is the obvious next study.

An option-shuffle intervention on EGOSCHEMA. The annotation-side half of this is done (the permutation null of Table 2 moves every gold to a uniformly random slot), and the model-side half is implemented in the same script as the leading-token map, against the pre-registered baseline position-bias measurement in `results/a2_interventions/baseline_position_bias.json`. It has not been run. Since §D.11 finds the larger half of EGOSCHEMA's excess to be positional, this is the intervention that would test the paper's own recommended repair directly.

Neither changes a claim the paper makes; each would extend one.

A third-party positive control, and what it calibrates. Every corpus in the table is third-party, but the one corpus whose defect was *known* (and therefore the only one on which the instrument's sensitivity could be calibrated rather than asserted) was the parent corpus of this programme, which we built. The control we wanted is a third-party corpus whose defect its own authors documented *and repaired*, so that one instrument can be run on both versions and the difference compared against the repair. NLVR2 (Suhr et al., 2019) supplies exactly that: its authors publish an UNBALANCED split and a BALANCED one built to remove a signal a model could exploit without the images. We run this paper's own estimator (leave-one-out modal accuracy inside template groups against the permutation null of §3, not a proxy) on 3,562 and 2,300 items respectively, with the sentence as the question, the true/false verdict as the answer, and the corpus's own `synset` label as the coarsest template. The annotation files and the result are kept beside every other input to this paper rather than in a scratch directory, so the control is reproducible.

The instrument separates them. At the corpus's own category granularity the blind bound is 43.9% on the unrepaired split against 0.0% on the repaired one, a gap of $+43.9$ points; one refinement step in, where neither split is degenerate, it is 25.1% against 15.3%, a gap of $+9.9$. The marginal answer rate moves by only $+0.6$ points between the splits, so what the instrument is registering is not the label balance: it is the template-conditional structure the repair was aimed at. That is a measurement of this instrument's sensitivity against a repair somebody else performed and documented.

Two properties of the control have to be stated with it. First, 0.0% at the coarsest granularity is a degeneracy and not a reading: a group whose two answers are present in exactly equal number makes the leave-one-out mode the *other* label for every item, so a perfectly balanced group scores zero rather than chance. The $g \geq 1$ figures are the ones to read. Second, the excess over the permutation null is negative on both splits (-6.2 and -50.0 at $g{=}0$; -23.5 and -32.4 at $g{=}1$), because a binary corpus engineered toward balance is *more* balanced than a shuffle of its own answers. That is the excess statistic correctly detecting deliberate balancing, and it is also the reason the excess column of Table 1 should not be read on binary corpora as though a negative value meant an absence of structure. The announcement count V and the blind bound β remain readable there; the excess does not.

WHAT WAS FIXED BEFORE THE DATA WERE SEEN, AND WHAT WAS NOT

The companion submissions in this programme each carry a registration note, and this one does not, because there was nothing to register in the sense those notes use: no numerical prediction was made in advance of a measurement here. What can be said honestly is which parts of the instrument were frozen before which corpora were read, and we say it rather than leave the absence to be inferred.

Fixed before any corpus was read, and unchanged since: the stopword list above; the announcement rule as verbatim containment of content tokens; the leave-one-out modal estimator as a single-valued rule with the fixed blind-visible tie-break stated in §3; the permutation null and the decision to report the excess rather than $\hat{\beta}$; the refinement sweep over $g \in \{0, \ldots, 8\}$ and the leading-token template family; and the five original option-regime rules of Definition 10. Raised after the review, and

therefore a stated amendment rather than a frozen choice: the permutation budget from $R = 20$ to $R = 1000$ with the convergence run in `results/a2_nullconv.json` showing that the $R = 20$ Monte Carlo error (0.56 points at g=0, 2.2× the $R = 20$ SE) falls below 0.05 points at $R = 1000$.

Fixed before the corpus it was applied to, which is the sense in which the extra corpora tested the instrument: everything in the previous paragraph, on the 7 rows added after the original two. Those rows are an out-of-sample run of the free-form estimator, the null and the refinement analysis.

Decided after seeing data, and therefore not a test of anything: the discriminating announcement count V^{disc} (Definition 12), which was written in response to VSI-BENCH scoring $V = 100.00\%$; the four added option-regime rules, which were chosen from the multiple-choice literature after two option corpora had been read; the split estimator, which was written after the refinement trap was measured; and the bootstrap diagnosis, which was written after the inflation was observed. Each can only raise a bound (Proposition 1(ii)) or is a diagnostic rather than a bound, so none of them can manufacture a condemnation; but none of them is a prediction confirmed either.

That last sentence is sufficient for a *bound* and is not sufficient for a *p-value*, and the distinction is what fixes how the option-regime test is reported. Four of the scored strategies were selected with two corpora in view, so a family-wise test over the expanded family prices in a selection that was partly made by us rather than by the null. The confirmatory test of Table 2 therefore maximises over the frozen five only, and every class-level statement in this paper reads that half; the expanded family's test is reported beside it as sensitivity. The two agree on all 5 corpora (0 verdicts differ), which is a fact about these corpora and not a licence to conflate the families: had they disagreed, the frozen half would still have been the one quoted.

Three further analyses were added after the results were seen and are robustness checks rather than tests: the mask-granularity sweep of Appendix G, which re-partitions predictions already collected and adds no run; the matched-dose intervention of §D.9, whose map was written after the first intervention's dose heterogeneity and its ungrammatical output on polar questions had both been measured, and which is reported as a second design beside the first rather than in place of it; and the per-corpus cost measurement of §C.8, which replaced a runtime claim the manuscript had been making inconsistently. None can change a bound and none is a prediction confirmed. The matched-dose map was fixed, and its dose homogeneity verified on the corpus, before either backbone was re-run.

Pre-specified in direction only: the positive control of §4.2. The defect and its mechanism were recorded in the parent paper's deviation log before this instrument was written, and the expected direction (the instrument fires before the repair and not after) was fixed by that record. No effect size was predicted. The NLVR2 control above was run after the parent-corpus control and its direction was fixed by NLVR2's authors' own account of what their balanced split removes; its magnitude was not predicted either.

STOPWORD LIST

The content-token filter removes: `the`, `a`, `an`, `of`, `in`, `on`, `at`, `to`, `is`, `are`, `was`, `were`, `this`, `that`, `there`, `it`, `and`, `or`, `with`, `for`, `from`, `by`, `as`, `be` and `been`. It is short on purpose. A longer list raises V by discarding more of the answer before the conjunction is tested, so the choice is in the direction that would flatter the result, and it is stated rather than tuned.

WHY THE PER-CATEGORY EXCESS IS REPORTED ONE REFINEMENT STEP IN

Within a single category at g=0 the partition has one group, and the permutation null reshuffles that group's own answer multiset, so the null equals the estimate and the excess is identically zero by construction rather than by measurement. The per-category table therefore reports the excess at g=1, and prints β at both granularities so the two are not confused.