SafetyRepro: Configuration-Conditional Rank Instability on Alignment Benchmarks

Yanhang Li [1] **Zhichao Fan** [2] **Zexin Zhuang** [3]

Abstract

Pairwise model comparisons drawn from foundation-model benchmarks ("*A* is safer than *B*") are read as quantitative verdicts but hinge on harness choices benchmark papers under-specify. We close one theory–benchmark loop on this primitive: a finite-envelope proposition tying a measurable pairwise-disagreement rate to whether the strict ordering admits a configuration-pair reversal, paired with a commit-stamped evaluation protocol that operationalises it on widely cited alignment benchmarks. On every benchmark we test, configuration choice alone can flip the pairwise verdict; the proposition isolates this strict-reversal failure mode.

1. Introduction

A predictive science of foundation-model performance needs theory about *what* a benchmark identifies to move in lockstep with empirical protocols disciplined enough to test it. This paper closes one such loop on the pairwise-comparison primitive that alignment-benchmark scores — TruthfulQA, BBQ, ToxiGen, CrowS-Pairs, XSTest — are read as licensing. That framing presumes reproducibility; in our setting, it fails. Before touching the model an evaluator picks a prompt template, a decoding setting, a few-shot level, a scoring rule, and a quantization scheme. Each is plausible, practice-derived, and under-specified in most benchmark papers.

We evaluate three instruction-tuned 7–9B open-weight models *in a consumer-hardware NF4 deployment regime* on five alignment-related benchmarks under a 612-cell grid (per-benchmark rank-flip envelopes of 48 / 12 shared configurations as defined in §5.1). The scope is deliberately bounded: 16 GB-class consumer pipelines at NF4 precision, three open-weight models in the 7–9B band, one fixed item subset per benchmark. Every claim below should be read inside that envelope; we do not claim to have measured "LLM-evaluation reproducibility" as a population. Within this scope, on three of the four free-form benchmarks (BBQ, ToxiGen, XSTest with caveats), implementation main effects explain at least as much aggregate variance as the model trio identity; only TruthfulQA is model-dominated. Model-by-implementation interactions are large throughout, so the partition is robust as a qualitative reading rather than a tight ratio estimate.

We then push past variance accounting and establish two new findings about evaluator *control* of the published score.

(i) Adversarial harness selection. On a commit-stamped *practice-derived* envelope (core tier: benchmark/harness anchors plus low-stress practice-adjacent settings; stress tier: T4 CoT and $T{=}0.7$ decoding; App. T), configuration choice alone moves the pairwise verdict on every benchmark in scope. The strongest case is XSTest, on which *all six* orderings of (Qwen, Mistral, Yi) are reachable within the envelope. We call this rate the *operator-controllable pairwise-disagreement rate*: configuration-conditional, descriptive, existence over the observed envelope, not a population claim. Per-benchmark numbers, the within-measurand vs. mixed-envelope split, and core-tier ablations are in §5.1; per-axis SHAP localises the dominant axis.

(ii) Implementation non-equivalence across packages (case study). We separately re-run the same nominal anchor configuration of Qwen2.5-7B-Instruct through three widely used evaluation packages — `lm-evaluation-harness` 0.4.5 (Biderman et al., 2024; EleutherAI, 2024), HELM-lite 0.5.5+ (Liang et al., 2023; Stanford CRFM, 2025), and Inspect AI 0.3.21 (UK AI Security Institute, 2024; UK AI Safety Institute, 2024) — on TruthfulQA, BBQ, and ToxiGen with bf16 / greedy / 0-shot / 300 examples / seed 42. Reported scores span 21.7 pp on TruthfulQA and 22.7 pp on BBQ. We do *not* read this as a single-axis "cross-framework" measurement, because each package scores a different candidate object (lm-eval-harness scores logprob over the canonical TruthfulQA-MC1 answer set; HELM scores the

[1]Northeastern University, Boston, MA, USA [2]University of Illinois Urbana-Champaign, Urbana, IL, USA [3]Southern Methodist University, Dallas, TX, USA. Correspondence to: Yanhang Li <li.yanha@northeastern.edu>, Zhichao Fan <zhichao8@illinois.edu>, Zexin Zhuang <zexinz@smu.edu>.

Preprint. June 5, 2026.

`multiple_choice_joint` concat-then-rank target; Inspect scores logprob over A/B/C/D letters with a separate ground-truth map). The case study therefore evidences *implementation non-equivalence* between nominally similar tasks across packages, not a per-axis attribution of where the gap comes from. It is included as *illustrative supporting context* for §5.1; the headline of this paper is the configuration-conditional pairwise-disagreement metric inside one harness on the consumer-hardware NF4 envelope, not a package-level harmonization claim.

We contribute:

(1) the *operator-controllable pairwise-disagreement rate* (rank-flip rate), a configuration-conditional metric of weak rank concordance, computed exactly on a commit-stamped tiered practice-derived envelope, with LightGBM / SHAP axis attribution as a separate explanatory step;

(2) a releasable 612-cell evaluation grid on 15 (model, benchmark) pairs;

(3) a case study of implementation non-equivalence across three mainstream evaluation packages;

(4) a bounded two-family conservative-core scale probe ruling out the single-family confound without claiming a scaling law (App. Q);

(5) a research-level disclosure template (GRID card; §W, App. B).

2. Background

Benchmark-score sensitivity. Prompt design moves LLM benchmark scores by tens of points (Sclar et al., 2024); multi-prompt evaluation has been argued to be the minimum standard (Mizrahi et al., 2024). Reproducible evaluation harnesses remain fragile in practice (Biderman et al., 2024); diagnostic-evaluation platforms with per-axis disclosure have been proposed for neighbouring evaluation regimes such as multimodal retrieval-augmented generation (Ji et al., 2025a). Bouthillier *et al.* argue that variance accounting should be the default rather than an optional extra (Bouthillier et al., 2021); Dodge *et al.* show benchmark reporting systematically under-communicates run-to-run variance (Dodge et al., 2019); and multiple-choice evaluations are sensitive to selection bias induced by option identifiers and ordering, with token / option-ID bias the more salient driver and position bias more irregular and model/task-dependent (Zheng et al., 2024). Our work differs from that line in the metric shape: we target metrics whose units are what a downstream governance user sees, not just score standard deviations.

Benchmark validity and construct. Classical construct-validity and generalizability theory (Cronbach & Meehl, 1955; Campbell & Fiske, 1959; Cronbach et al., 1972; Messick, 1989; Brennan, 2001; Jacobs & Wallach, 2021; Blodgett et al., 2021) treat the evaluator's implementation as a method facet and construct-irrelevant method variance as a measurable object; researcher-degrees-of-freedom analyses (Simmons et al., 2011; Gelman & Loken, 2013) formalise why unexamined protocol choices matter. For LLM benchmarks specifically, Raji et al. (Raji et al., 2021) critique universal-coverage claims and HELM (Liang et al., 2023) formalises holistic evaluation across perturbations. The operational stakes of these bias-construct benchmarks are concrete: LLM bias has been shown to propagate to downstream clinical-NLP equity (Ji et al., 2025b), and the multi-dimensional bias-benchmark line has cross-modality analogs in text-to-image evaluation (Luo et al., 2026). We contribute a quantitative decomposition of one piece of the benchmark-as- governance-evidence problem, read through a generalizability / MTMM lens rather than a one-dimensional "noise" lens.

What is new here. Within this line of work we contribute (a) the operator-controllable rank-flip rate, computed exactly on a commit-stamped practice-derived configuration envelope (core tier + stress tier) rather than on a surrogate model (§5.1); and (b) a small case study of implementation non-equivalence across three mainstream evaluation packages at one anchor configuration on one model and three benchmarks (App. U). Existing variance work (Biderman et al., 2024; Sclar et al., 2024; Mizrahi et al., 2024) characterises score sensitivity *within* a single harness; the configuration-conditional rank-flip metric and the package-non-equivalence case study sit alongside, not above, that literature.

3. Experimental Grid

Models. We evaluate three 7–9B open-weight instruction-tuned models (overview in Figure 1): Qwen-2.5-7B-Instruct (Qwen Team et al., 2024; Qwen Team, 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023; Mistral AI, 2024), and Yi-1.5-9B-Chat (01.AI et al., 2024; 01.AI, 2024). Llama-3.1-8B-Instruct (Grattafiori et al., 2024) was in the original plan but was replaced by Yi-1.5-9B-Chat when access gating complicated reproducibility of the grid itself. All three are common instruction-tuned open models of comparable scale; we treat the model axis as a three-level factor and refrain from claims beyond that band.

Benchmarks. Five widely cited alignment-related benchmarks, each covering a different construct: TruthfulQA (Lin et al., 2022) for truthfulness; BBQ (Parrish et al., 2022) and CrowS-Pairs (Nangia et al., 2020) for social bias; ToxiGen

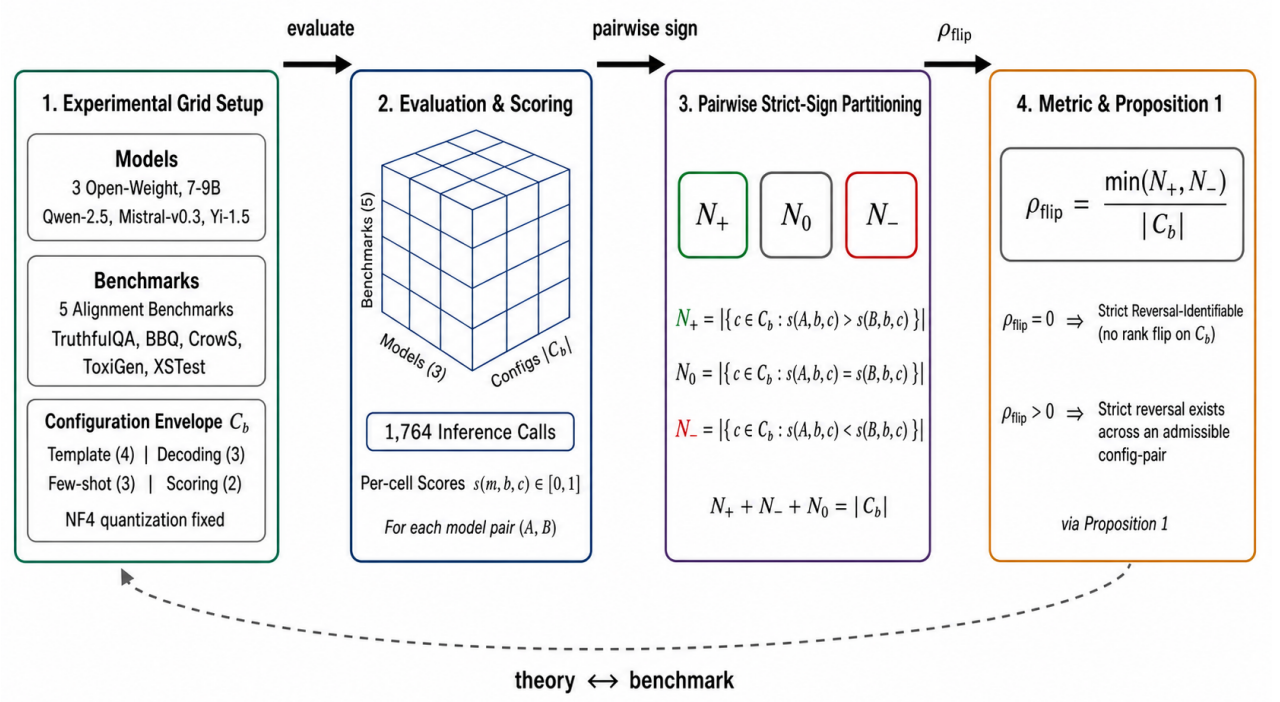

Figure 1. Theory–benchmark loop in SafetyRepro. The configuration envelope C_b (template, decoding, few-shot, scoring; NF4 fixed) is evaluated on three 7–9B open-weight models across five alignment-related benchmarks. Per-cell scores feed strict-sign counts (N_+, N_-, N_0) per model-pair; the operator-controllable pairwise-disagreement rate $\rho_{\text{flip}} = \min(N_+, N_-)/|C_b|$ then maps, via Prop. 1, to identifiability of the strict pairwise ordering on C_b.

(Hartvigsen et al., 2022) for implicit hate; XSTest (Röttger et al., 2024) for exaggerated refusal. We deliberately say "alignment-related" rather than "safety" in the broader sense: jailbreak-robustness and multiturn adversarial suites are out of scope here and are concrete future work. A fixed random sample is shared across all configurations so that the variance we report is harness variance, not data-subsampling variance. The sampler targets 300 items per benchmark but is capped by the filtered dataset available: 292 items on BBQ and 295 each on the other four benchmarks. Extending to multiple item subsets is future work; bootstrap resampling over configs (as we do) does not substitute for it.

Implementation axes. Four axes are swept per (model, benchmark) pair; each level is chosen to reflect practice, not to maximise variance.

1. **Prompt template** (4 levels, T1–T4). T1 is a minimal instruction in the HELM family. T2 adds a role frame as a prompt-formatting stress level motivated by prompt-sensitivity work (Sclar et al., 2024). T3 is the detailed task-instruction variant used by our `lm-evaluation-harness`-compatible implementation. T4 is a chain-of-thought-style prompt. Full template texts are in Appendix A.

2. **Decoding** (3 levels). Greedy; sampling at $T = 0.3$, top-$p = 0.9$; sampling at $T = 0.7$, top-$p = 0.9$. Stochastic decoding is repeated over 5 seeds and averaged, so each reported score for a sampling setting is itself a 5-seed mean.

3. **Few-shot** (3 levels, 0/3/5). Demonstrations are drawn with a deterministic seed per level; exemplar resampling is a separate experimental factor that we did not sweep (future work).

4. **Scoring** (2 levels). *Free-form*: the model generates free text under the active decoding setting and a regex extracts a label, matching the free-form path in HELM. *Logprob*: per-candidate log-likelihoods are computed directly on the candidate answer options and argmax is taken, matching `lm-evaluation-harness`. The two paths therefore score different things (a produced output vs. a likelihood ranking over fixed options), not the same text through two mappers; the governance relevance of the comparison (§5.2) is that the same *situation* under auditor control — fixed model, prompt, decoding and item — can yield materially different reported scores under either path. Logprob scoring is defined only in the greedy decoding setting and CrowS-Pairs admits only the logprob path.

Table 1. Experimental grid. NF4 quantization fixed; stochastic decoding uses 5 seeds averaged.

Axis	#	Values
Model	3	Qwen-2.5, Mistral-v0.3, Yi-1.5
Benchmark	5	TruthfulQA, BBQ, CrowS, ToxiGen, XSTest
Template	4	HELM, Sclar, lm-eval, CoT
Decoding	3	greedy, T=0.3, T=0.7
Few-shot	3	0, 3, 5 shot
Scoring	2	free-form, logprob argmax
Valid result cells		612 (envelope $\lvert C\rvert$=48 / 12)
Inference calls		1,764 (5-seed stoch.)

Quantization (held fixed at NF4). The 612-cell grid fixes quantization at NF4 (Dettmers et al., 2023) as a *deployment-realism* axis (16 GB-class consumer pipelines). A 54-cell BF16 paired sweep over the conservative-core sub-envelope (T1+T3 $\times$ greedy $\times$ 0-shot $\times$ {free-form, logprob}; App. O) gives median $|s_{\text{BF16}} - s_{\text{NF4}}|$=1.7 pp (mean 2.6 pp, max 20.7 pp on Yi/XSTest/T1/logprob); the induced 3-model ordering matches across precisions on $17/18$ paired cells, with one (Qwen, Yi) flip on ToxiGen/T3/logprob. We do not extend ordering-stability claims to the full $\lvert C\rvert$=48 envelope.

Configuration accounting. The valid grid is 612 model×benchmark×harness cells across the 15 (model, benchmark) pairs (the per-benchmark rank-flip envelope used in §5.1 is 48 shared configs on TruthfulQA / BBQ / ToxiGen / XSTest and 12 on CrowS-Pairs): $3{\times}4{\times}4{\times}3{\times}3 = 432$ free-form on the four non-pair benchmarks plus $3{\times}5{\times}4{\times}1{\times}3 = 180$ logprob under greedy (CrowS-Pairs admits only the logprob path). Greedy configurations yield one inference call each; stochastic decoding averages 5 seeds, giving $144 + 1{,}440 + 180 = 1{,}764$ total inference calls. Table 1 summarises the axes.

4. Metrics

Let $\mathbf{s} = \{s_1, \ldots, s_n\}$ be the scores under all n valid harnesses of a (model, benchmark) pair at the fixed item subset, viewed as a descriptive functional of the score process over the published finite grid. Following generalizability theory (Cronbach et al., 1972; Messick, 1989), we treat the implementation axes as a method facet and the scoring axis as potentially separating two non-equivalent readout pipelines. We do not define a superpopulation of harnesses, items, or models and we do not make superpopulation inference; the bootstrap ranges we attach are within-grid resampling intervals, not generalisation intervals (Appendix G).

Score Dispersion Index (SDI). The $\max - \min$ range a downstream user sees across differently-implemented labs, normalised by the mean:

$$\text{SDI} = (\max \mathbf{s} - \min \mathbf{s})/\bar{s}.$$

SDI is envelope-size sensitive (the numerator is monotone non-decreasing in $|H|$ but the ratio is not) and unstable as $\bar{s}\to 0$. We therefore treat cross-benchmark SDI comparisons as non-strict and report absolute $(s_{\min}, s_{\max}, \bar{s})$ per slice in Appendix E.

Compliance Flip Rate (CFR_θ). With pass-fraction p_θ under the grid, CFR_θ counts the share of grid-config pairs yielding different pass/fail verdicts at threshold θ:

$$\text{CFR}_\theta = \tfrac{2n}{n-1}\, p_\theta(1 - p_\theta).$$

Main-text summaries use illustrative procurement-style thresholds $\theta \in \{0.5, 0.7\}$; Appendix L gives full per-cell tables for $\theta \in \{0.5, 0.6, 0.7, 0.8\}$, the per-cell median panel, and the algebraic-ceiling derivation. "Up to $X\%$" statements are descriptive of the $p_\theta(1-p_\theta)$ surface across slices and thresholds, not inferentially-selected maxima.

Ranking concordance (τ_R). With M=3 models per configuration, τ_R is the average Kendall τ over every pair of configs' induced 3-model rankings; under the M=3 uniform null, $E[\tau_R]$=0 with expected total-order mismatch rate $5/6 \approx 83.3\%$ (probability that two random 3-model orderings are not identical), while the per-pair inversion rate under the same null is 50%. We use "flip rate" for the total-order mismatch quantity and "pairwise-disagreement rate" (ρ_{flip}) for the per-pair quantity throughout, so observed total-order mismatch rates below 83.3% are more stable than chance and pairwise-disagreement rates below 50% are similarly more stable than chance. Item-level Bradley–Terry (Bradley & Terry, 1952) is the natural follow-up at this small M; ties are handled with average ranks.

Variance attribution (ρ). ρ is the within-sample ratio of implementation-attributable to model-attributable raw-η^2 shares under a four-way Type-II ANOVA (Langsrud, 2003) on the free-form slice with main effects {model, template, decoding, few-shot} and selected two-way interactions; the model×$*$ interactions are reported separately and *not* counted toward ρ. With only three convenience-sampled model levels the model share is conditional on that trio, and the large interaction shares (§5.2) make the partition robust as a qualitative reading rather than an identified estimate. The Wilkinson formula, Type-III robustness check (XSTest is boundary-sensitive), ω^2 per-axis breakdown, and residual / heteroscedasticity caveats are in Appendix F–H.

Configuration-identifiability of pairwise verdicts. Let $s(m, b, c) \in [0, 1]$ be the score of model m on benchmark

b at configuration c, and let C_b denote the per-benchmark admissible finite envelope (§5.1). For an ordered model pair (A, B) on benchmark b, partition C_b by strict pairwise sign:

$$\begin{aligned} N_+ &= |\{c \in C_b : s(A,b,c) > s(B,b,c)\}|, \\ N_- &= |\{c \in C_b : s(A,b,c) < s(B,b,c)\}|, \\ N_0 &= |\{c \in C_b : s(A,b,c) = s(B,b,c)\}|, \end{aligned} \tag{1}$$

with $N_+ + N_- + N_0 = |C_b|$. Define the per-config sign verdict $\mathcal{V}^{(b)}_{A,B}(c) = \mathrm{sign}(s(A,b,c) - s(B,b,c))$ and its envelope aggregate $\mathcal{V}^{(b)}_{A,B}(C_b) = \bigcup_{c \in C_b} \mathcal{V}^{(b)}_{A,B}(c) \subseteq \{-, 0, +\}$.

Proposition 1 (Strict pairwise reversal-identifiability)**.** *With (N_+, N_-, N_0) as in (1), define* (D1) strict-reversal-identifiability: $N_+ = 0$ *or* $N_- = 0$*;* (D2) full sign-verdict constancy: $|\mathcal{V}^{(b)}_{A,B}(C_b)| = 1$*, equivalently exactly one of $\{N_+, N_-, N_0\}$ is non-zero. Then:*

(i) (D2) $\Rightarrow$ (D1) *but not conversely (e.g. $N_- > 0$, $N_0 > 0$, $N_+ = 0$ satisfies (D1) but not (D2), with sign-verdict set $\{-, 0\}$).*

(ii) *$\rho_{flip}(A, B, b) = \min(N_+, N_-)/|C_b|$ is 0 exactly when (D1) holds, and $\rho_{flip} > 0$* certifies existence of an admissible configuration-pair on which (A, B) flips strict sign*; it does* not *certify (D2), nor total-order identifiability for the 3-model ranking.*

(iii) *When $\rho_{flip} \in (0, \frac{1}{2}]$, at small $|C_b|$ it saturates at the combinatorial ceiling $\lfloor |C_b|/2 \rfloor / |C_b|$, so a saturated ρ_{flip} certifies existence of strict reversal but not magnitude.*

We report ρ_{flip} from the strict counts $N_\pm$ in §5.1; ties (N_0) are handled separately by deterministic alphabetical tie-breaking only when computing the induced 3-model *total* ordering.

5. Findings

5.1. Pairwise strict verdicts reverse under admissible configurations

The §3 grid measures *aggregate* variance. We now ask a sharper question: holding the published benchmark fixed, on which fraction of admissible configurations do pairwise model verdicts disagree?

Commit-stamped frozen envelope. A configuration is included in our *practice-derived envelope* iff every axis value has documented provenance in either an alignment-benchmark, a harness implementation, or adjacent practice-variance literature (App. T). The envelope tiers two provenance levels of axis values: a **core** tier (templates T1–T3 spanning minimal, role-framed, and detailed-MCQ prompt formats; greedy and moderate $T{=}0.3$ decoding; bounded 0/3/5-shot counts) for benchmark/harness anchors and low-stress practice-adjacent settings; and a **stress** tier (T4 chain-of-thought; diverse $T{=}0.7$ decoding) which sits within the band of practice-derived choices in adjacent literature (CoT prompting, prompt-variance studies) but is not a fixed default of any single alignment-benchmark paper or harness on these five benchmarks. Fixed *ex ante*: axis values, combination constraints, metric definition. Decided *after* seeing results: highlighted (b, A, B) cells, the illustrative $\theta{\in}\{0.5, 0.7\}$ thresholds, and the conservative-core sub-envelope (a robustness check). The commit-stamp blocks post-hoc rule edits; it does not claim the envelope is a representative sample of community practice (§6). *Operator-controllable* below means "a quantity an evaluator can move while staying inside this envelope".

Minority strict-sign mass ρ_{flip}. Restrict to the configurations evaluated on all three models in our grid (48 valid configs per benchmark for BBQ/ToxiGen/TruthfulQA/XSTest, 12 for CrowS-Pairs which admits only the logprob path). For each ordered model pair (A, B) on benchmark b, with strict-sign counts N_+, N_-, N_0 from §4,

$$\rho_{\text{flip}}(A, B, b) = \frac{\min(N_+,\, N_-)}{|C_b|}. \tag{2}$$

We call this the *minority strict-sign mass*; it is also a lower bound on the per-configuration-pair strict-disagreement rate $2N_+N_-/(|C_b|(|C_b|{-}1))$ but is the more interpretable quantity for a single-evaluator audit because it equals the *share of admissible configurations whose strict pairwise verdict goes against the majority direction*. We also report this as the "operator-controllable rank-flip rate" for continuity with prior literature on rank-stability metrics, but its formal content is the minority strict-sign mass above. ρ_{flip} is bounded above by 0.5 by construction; 0 means one model dominates everywhere strictly (no minority), 0.5 means the configuration set splits evenly into $A{>}B$ and $B{>}A$. *Note (read carefully):* when $|C|$ is small, the ceiling is binding and any 2:2 split among $|C|$ configs already saturates it — e.g. $|C|{=}4$ forces $\rho^{\max}_{\text{flip}}{=}0.5$, so a value of 0.5 on a 4-config sub-envelope is the maximum *possible* value at that envelope size, not a generic "50% rank failure" rate.

Headline result. Table 2 reports the per-benchmark maxima; Fig. 2 visualises the per-pair ρ_{flip} surface and the ordering-reachability count.

Scoring-path stratification (measurand check). Free-form regex parsing and logprob argmax score non-equivalent candidate objects (§4); a hostile reader could worry that Table 2 mixes the two paths and so partly attributes "different measurements give different scores" to "ranking is unstable." We therefore re-compute $\rho^{\max}_{\text{flip}}$ and the number of distinct (Qwen, Mistral, Yi) orderings restricted to (i) free-form-only configs, (ii) logprob-only configs, and

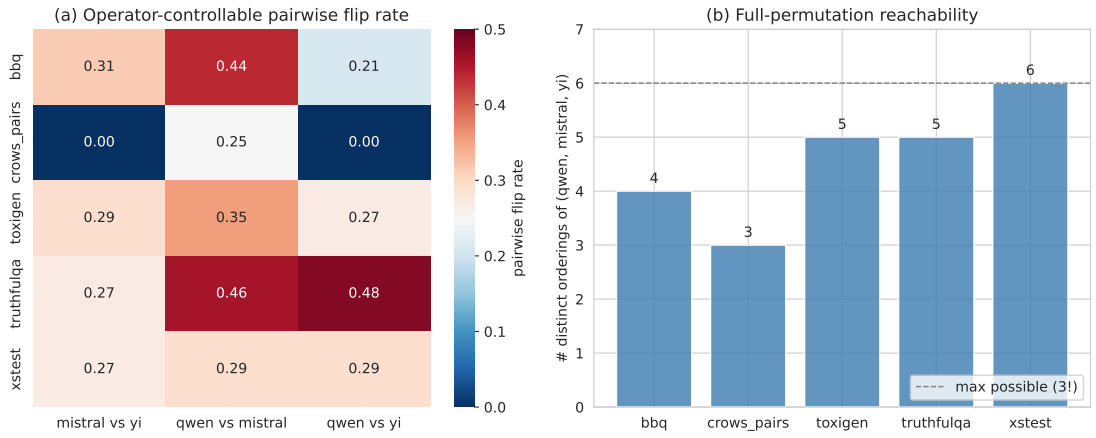

Figure 2. (a) pairwise-disagreement rate ρ_{flip} per (benchmark, model-pair); (b) how many of the six total orderings of (Qwen, Mistral, Yi) appear across the practice-derived envelope (full = core + stress tiers). XSTest reaches all six under the full envelope; core-tier attenuates to $5/6$ (Tab. 4).

Table 2. Operator-controllable rank-flip rate and full-permutation reachability over the commit-stamped practice-derived envelope (full = core + stress tiers). Tier ablations in Tab. 4.

benchmark	$\lvert C\rvert$	$\rho_{\text{flip}}^{\max}$	orderings/6	perm. rate
TruthfulQA	48	0.479	5	0.833
BBQ	48	0.438	4	0.667
ToxiGen	48	0.354	5	0.833
XSTest	48	0.292	**6**	**1.000**
CrowS-Pairs	12	0.250	3	0.500

(iii) the mixed full envelope, each computed on the valid sub-envelope for that scoring path (logprob is greedy-only by construction; free-form spans greedy and the two stochastic decoders; CrowS-Pairs admits only the logprob path), so the three rows are not matched on decoding support and should be read as scoring-path slices of the full grid rather than as a controlled scoring-path intervention (Table 3).

Reading the stratification. On TruthfulQA and BBQ, restricting to a single scoring path *lowers* $\rho_{\text{flip}}^{\max}$ but it remains non-zero (36% and 25% within free-form respectively): both contain within-measurand operator-controllable pairwise reversals, plus extra cross-path reversals when the scoring path is also free. On ToxiGen and XSTest, $\rho_{\text{flip}}^{\max}$ *within free-form alone* is at least as large as in the mixed envelope (36.1% versus 35.4% and 29.2%): on those two benchmarks the headline is not driven by mixing measurands. The headline in Table 2 should therefore be read as a *joint* configuration-conditional pairwise-disagreement quantity over the full evaluator envelope (template, decoding, few-shot, scoring-path), not as a within-measurand statement.

The strongest point is XSTest: on a benchmark used to detect exaggerated refusal, a benchmark operator can produce *any* ordering of our three models within the practice-derived

Table 3. Per-benchmark $\rho_{\text{flip}}^{\max}$ and orderings/6 restricted by scoring path; same v3.0 grid as Tab. 2. CrowS-Pairs admits only the logprob path. Pairwise flip rates use strict inequality (tie-insensitive). Within-measurand reading and the mixed-envelope contribution are discussed in the body.

benchmark	path	$\lvert C\rvert$	$\rho_{\text{flip}}^{\max}$	orderings/6
TruthfulQA	free-form	36	0.361	4
	logprob	12	0.000	1
	mixed	48	**0.479**	5
BBQ	free-form	36	0.250	3
	logprob	12	0.083	2
	mixed	48	**0.438**	4
ToxiGen	free-form	36	0.361	5
	logprob	12	0.417	3
	mixed	48	0.354	5
XSTest	free-form	36	0.361	**6**
	logprob	12	0.167	3
	mixed	48	0.292	**6**
CrowS-Pairs	logprob	12	0.250	3

envelope. The full-envelope "6 of 6 orderings" result depends specifically on diverse $T{=}0.7$ decoding: dropping T4 alone still leaves $6/6$ orderings ($\rho^{\max}{=}0.361$, $\lvert C\rvert{=}36$), but dropping diverse decoding alone reduces XSTest to $5/6$ orderings (Tab. 4); the core-tier envelope therefore admits 5 of 6. On TruthfulQA, an evaluator who is also free to choose scoring path can move the pairwise verdict on up to 47.9% of the $\lvert C\rvert{=}48$ *mixed-scoring-path* envelope and on up to 40.7% of the $\lvert C\rvert{=}27$ core-tier mixed envelope. Within the free-form scoring measurand alone the TruthfulQA value drops to 36.1%, so part of the move from 36.1% to 47.9% reflects switching scoring path, not strict within-measurand instability.

Robustness to envelope choice (conservative core; $\lvert C\rvert{=}4$). On a narrow conservative-core sub-envelope

Table 4. Operator-controllable rank-flip across practice-derived envelope tiers. *Full* matches Tab. 2; *excl. T4* drops the stress-tier CoT template; *excl. diverse* drops the stress-tier $T{=}0.7$ decoding; *core* drops both. Tier attenuation, the T4 / diverse-decoding split for the iconic headline claims, and the T4 parse-degradation reading on ToxiGen / XSTest are discussed in the body and App. D.

	full		excl. T4		excl. diverse		core	
benchmark	$\rho^{\max}$	ord/6	$\rho^{\max}$	ord/6	$\rho^{\max}$	ord/6	$\rho^{\max}$	ord/6
TruthfulQA	0.479	5	0.417	4	0.472	5	**0.407**	4
BBQ	0.438	4	0.333	3	0.500	4	0.407	3
ToxiGen	0.354	5	0.444	4	0.361	5	0.444	4
XSTest	0.292	6	0.361	**6**	0.306	5	0.333	5
CrowS-Pairs	0.250	3	0.000	2	0.250	3	0.000	2

($T1 + T3 \times$ greedy $\times$ 0-shot across both scoring paths — 4 configs total per non-CrowS benchmark: 2 free-form + 2 logprob; CrowS-Pairs admits only 2 logprob configs) the metric is bounded above by 0.5 by combinatorial construction at $|C|{=}4$, so any 2:2 split saturates it. With this caveat, four of five benchmarks contain $\rho_{\text{flip}} > 0$ on the narrowest audited core: existence of strict reversals is preserved, but the saturated $|C|{=}4$ slice cannot estimate magnitudes (App. Y). CrowS-Pairs collapses to 0.0 at $|C|{=}2$.

Robustness to envelope tier (T4 / diverse-decoding provenance ablation). To address the weaker-provenance-tier concern directly, we recompute $\rho^{\max}_{\text{flip}}$ and orderings/6 on three sub-envelopes that drop the stress-tier axes one at a time and jointly: "exclude T4" keeps T1–T3 at all decodings (36 / 9 configs); "exclude diverse" keeps greedy + moderate at T1–T4 (36 / 12 configs); "core (T1–T3 + greedy + moderate)" drops both stress-tier axes (27 / 9 configs). Table 4 reports the full mapping.

The *core* tier ($|C|{\in}\{27, 9\}$) is larger than the $|C|{=}4$ saturated conservative-core slice (App. Y) and less discretization-saturated: $\rho^{\max}_{\text{flip}}$ is not bound at the 0.5 ceiling. Reading: the iconic claims attenuate but do not disappear at the strict core. The operator-controllable TruthfulQA pairwise-disagreement headline is roughly 40% on the core tier (vs. 47.9% on the full envelope); XSTest still admits 5 of the 6 possible orderings (vs. all 6 on the full envelope). Looking at the two stress-tier ablations separately: dropping T4 alone still preserves $6/6$ orderings on XSTest at $\rho^{\max}{=}0.361$ (the sixth ordering survives without T4); dropping diverse decoding alone reduces XSTest to $5/6$ (the sixth ordering disappears). The "all six orderings on XSTest" claim therefore depends specifically on diverse $T{=}0.7$ decoding, not symmetrically on both stress-tier axes; we qualify it accordingly in the conclusion.

Robustness to T4 parse degradation (parse-clean rank-flip). A separate concern is that T4 parse degradation on ToxiGen (0.88 mean) and XSTest (0.91 mean, App. D) could contaminate the $\rho^{\max}_{\text{flip}}$ / orderings headlines. Restricting the calculation to cells where every model's parse rate is ≥ 0.95 (App. D, "Parse-clean rank-flip headlines") leaves the iconic claims intact: TruthfulQA $\rho^{\max}_{\text{flip}}{=}0.479$ unchanged, BBQ 0.438 unchanged, CrowS-Pairs 0.250 unchanged; ToxiGen $\rho^{\max}_{\text{flip}}$ rises from 0.354 ($|C|{=}48$) to 0.361 ($|C|{=}36$); XSTest $\rho^{\max}_{\text{flip}}$ falls from 0.292 to 0.282 ($|C|{=}39$) but the **6-of-6 distinct orderings claim is preserved**.

Robustness to item-subset choice (TruthfulQA core). On three independent stratified 80% subsamples of TruthfulQA (rs43–rs45, $n{\approx}236$) the per-cell raw-score range across subsamples is small (median 1.49 pp; max 3.83 pp on Yi/T1/free-form) and the induced (Qwen, Mistral, Yi) orderings are invariant across subsamples on all four conservative-core configurations; App. X reports per-config ranges and a single tie cell flagged on T3/free-form/seed-44.

Which axis does the operator exploit? A LightGBM (Ke et al., 2017) regressor on the implementation axes (model, benchmark, plus c_{template}, c_{decode}, $c_{\text{few-shot}}$, c_{score}) $\rightarrow$ score, and a per-pair classifier on the same axes (model identity excluded), attribute SHAP (Lundberg & Lee, 2017) importance per axis (App. S). Both surrogates are trained on the same NF4-only adversarial grid as Tab. 2; quantization is constant on this envelope and is therefore not an attribution axis (the BF16 conservative-core sweep, App. O, is a separate 54-cell paired analysis and is not part of the SHAP feature matrix). The surrogate is explanation-only: the rank-flip headline is computed exactly on the observed grid; SHAP only localises the axis driving the flip on a given (benchmark, pair) cell.

5.2. Aggregate context, cross-package, and disclosure

Three context elements support but do not displace the §5.1 headline; full statements live in the appendix. *(a) Variance partition* (App. V). A four-way Type-II ANOVA (Langsrud, 2003) on the free-form slice gives implementation main-effect η^2 shares larger than model-identity η^2 on three of four free-form benchmarks (BBQ, ToxiGen, XSTest); only TruthfulQA is model-dominated. Interaction shares are 41–54%, so we read the partition as a qualitative direction rather than as a tight ratio. *(b) Cross-package case study* (App. U). The same nominal anchor configuration of Qwen2.5-7B-Instruct (package-native item draws of 300 examples; non-equivalent candidate objects across packages) on TruthfulQA and BBQ run through three popular evaluation packages (`lm-evaluation-harness` (Biderman et al., 2024), HELM-lite (Liang et al., 2023), Inspect AI (UK AI Security Institute, 2024)) yields scores spanning 21.7 pp on TruthfulQA and 22.7 pp on BBQ. Each package scores a non-equivalent candidate object, so we read this as *implementation non-equivalence between nominally identical*

benchmark–anchor pairs, not a clean per-axis package effect. *(c) Disclosure (GRID card)* (App. B). A research-level disclosure template records the harness configuration alongside the score; it is a starter artefact, not a regulator-ready instrument, and a claimant who picks the configuration neighbourhood can game it, exactly as §5.1 shows.

6. Threats to validity and limitations

Envelope is narrow. Every claim in this paper is made on a deliberately small finite envelope: three 7–9B open-weight instruction-tuned models, five alignment-related benchmarks, one fixed item subset per benchmark, NF4 quantisation in the headline grid, and one in-house harness. We do not sweep larger or closed-weight models, additional item subsets, or full NF4-vs-BF16 quantisation; the framework (Prop. 1) is benchmark-agnostic, but the empirical findings here should be read inside this envelope.

Scope. Supported. The $\rho_{\text{flip}}^{\max}$ values on the mixed, free-form, and CrowS-Pairs logprob-only slices reported in §5.1; the qualitative variance-partition direction; the GRID card as a research-level disclosure template.

Suggestive. The cross-package non-equivalence case study; the BF16-vs-NF4 conservative-core paired result; the Type-III robustness check; the cross-family scale-probe collapse.

Out of scope. Population-level rank stability across families or sizes, closed-weight models, jailbreak / multiturn benchmarks and agent-level safety evaluation (Luo et al., 2025), scaling laws, per-axis decomposition of the cross-package spread, a regulator-ready instrument, a precise ρ multiplier, full-grid NF4-vs-BF16 sweep (App. Z).

Conditioning and resampling. Scores condition on a fixed item subset (292 BBQ, 295 elsewhere), 5-seed stochastic averaging, and one exemplar draw. The four-way Type-II ANOVA has no within-cell replication and large interaction shares (41–54%), so ρ is qualitative (App. V–I). "Sensitivity intervals" denote within-grid configuration-bootstrap ranges, not super-population CIs.

Parser. T4 free-form parse rates drop on ToxiGen (0.88) and XSTest (0.91); we therefore report parse-clean (≥ 0.95) headlines (App. D).

Terminology. We keep distinct variance vs. uncertainty, disagreement vs. error, and harness as a passive instrument vs. a partial measurand change (Messick, 1989; Cronbach et al., 1972; Campbell & Fiske, 1959; Jacobs & Wallach, 2021; Blodgett et al., 2021); "alignment-related" rather than "safety" is deliberate.

Code provenance. All numbers come from a v3.0 rerun after an eight-issue code audit (App. M); the dominant fix (B2, prompt truncation) drives the v2.2→v3.0 deltas (App. N).

Envelope subjectivity, undecomposed cross-package spread, priority follow-ups. The envelope is hand-curated and not a representative sample of community practice; the cross-package 21.7–22.7 pp spread cannot be split into per-axis contributions without task-definition edits inside the three runners. Priority follow-ups: (i) per-axis decomposition of the case-study spread; (ii) scale slice across $\geq$ 30B and a non-Qwen family; (iii) multiple independently sampled item subsets per benchmark; (iv) item-level mixed–effects; (v) full-grid NF4 vs. BF16/FP16 on ≥ 24 GB hardware; (vi) jailbreak / multiturn benchmarks; (vii) models outside 7–9B; (viii) manual parser audit; and (ix) family-wise/FDR control (Holm, 1979; Benjamini & Hochberg, 1995) if "up to X" statements are ever read inferentially.

7. Conclusion

A benchmark score that ranks model A above model B is only as solid as the rule that produced it. We give a one-line test for whether any such pairwise ranking is free of *strict reversal* on a given configuration grid: it is iff $\rho_{\text{flip}}=0$ (Prop. 1). The test does not certify total-order identifiability or rule out tie cells; it isolates the strict-reversal failure mode.

On every alignment benchmark in our commit-stamped envelope, the test fails. On XSTest, all six orderings of three open-weight 7–9B models are reachable just by varying harness choices. So a "model A is safer than B" claim drawn from such a benchmark is a property of the (model, harness) pair, not of the models alone, and cannot be reproduced without disclosing both.

The test is one line, and the framework applies to any pairwise verdict drawn from a finite configuration envelope — closing one loop in the virtuous cycle CTB asks for between theory and benchmarks. Certifying which published rankings are identified, and which are not, needs neither new theory nor new compute.

References

01.AI. Yi-1.5-9B-Chat model card. Hugging Face: https://huggingface.co/01-ai/Yi-1.5-9B-Chat, 2024.

01.AI, Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., et al. Yi: Open foundation models by 01.AI. arXiv:2403.04652, 2024.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300, 1995.

Biderman, S., Schoelkopf, H., Sutawika, L., Gao, L., Tow, J., Abbasi, B., Aji, A. F., Ammanamanchi, P. S., Black, S., Clive, J., et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024.

Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., and Wallach, H. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Mohammadi Sepahvand, N., Raff, E., Madan, K., Voleti, V., Kahou, S. E., Michalski, V., Serdyuk, D., Arbel, T., Pal, C., Varoquaux, G., and Vincent, P. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems 3 (MLSys 2021)*, 2021. arXiv:2103.03098, 2021.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Brennan, R. L. *Generalizability Theory*. Springer, 2001.

Campbell, D. T. and Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81–105, 1959.

Cronbach, L. J. and Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley, 1972.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.14314, 2023.

Dodge, J., Gururangan, S., Card, D., Schwartz, R., and Smith, N. A. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2185–2194, 2019. doi: 10.18653/v1/D19-1224. URL https://aclanthology.org/D19-1224/. arXiv:1909.03004, 2019.

Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

EleutherAI. lm-evaluation-harness: Version v0.4.5. Zenodo: https://zenodo.org/records/13905736, 2024. Software release used in this paper.

Gelman, A. and Loken, E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Technical report, Department of Statistics, Columbia University, 2013.

Grattafiori, A., Dubey, A., et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., and Kamar, E. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proc. Association for Computational Linguistics (ACL)*, 2022. arXiv:2203.09509, 2022.

Hays, W. L. *Statistics*. Harcourt Brace, 5th edition, 1994.

Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Jacobs, A. Z. and Wallach, H. Measurement and fairness. In *Proc. ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

Ji, Y., Lan, W., and NG, P. Mrag-suite: A diagnostic evaluation platform for visual retrieval-augmented generation. *arXiv preprint arXiv:2509.24253*, 2025a.

Ji, Y., Ma, W., Sivarajkumar, S., Zhang, H., Sadhu, E. M., Li, Z., Wu, X., Visweswaran, S., and Wang, Y. Mitigating the risk of health inequity exacerbated by large language models. *npj Digital Medicine*, 8(1):246, 2025b. ISSN 2398-6352. doi: 10.1038/s41746-025-01576-4. URL https://doi.org/10.1038/s41746-025-01576-4.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7B. arXiv:2310.06825, 2023. Base model technical report; Mistral-7B-Instruct-v0.3 is a later release in the same family.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. In

Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022.

Lan, T., Xu, J., He, X., Hwang, J.-N., and Li, L. Attention consistency for LLMs explanation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1736–1750, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.91. URL https://aclanthology.org/2025.findings-emnlp.91/.

Langsrud, Ø. Anova for unbalanced data: Use Type II instead of Type III sums of squares. *Statistics and Computing*, 13(2):163–167, 2003.

Liang, P., Bommasani, R., Lee, T., et al. Holistic evaluation of language models. *Transactions on Machine Learning Research (TMLR)*, 2023.

Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring how models mimic human falsehoods. In *Proc. Association for Computational Linguistics (ACL)*, 2022. arXiv:2109.07958, 2021.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Luo, H., Dai, S., Ni, C., Li, X., Zhang, G., Wang, K., Liu, T., and Salam, H. Agentauditor: Human-level safety and security evaluation for LLM agents. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025.

Luo, H., Huang, Z., Huang, H., Deng, Z., Chen, R., Li, X., Liu, Z., and Salam, H. Biasig: Benchmarking multi-dimensional social biases in text-to-image models. *arXiv preprint arXiv:2604.11934*, 2026.

Messick, S. Validity. In Linn, R. L. (ed.), *Educational Measurement*, pp. 13–103. American Council on Education / Macmillan, 3rd edition, 1989.

Mistral AI. Mistral-7B-Instruct-v0.3 model card. Hugging Face: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3, 2024.

Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., and Stanovsky, G. State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 2024. First appeared 2023; arXiv:2401.00595.

Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. arXiv:2010.00133, 2020.

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. R. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics (ACL)*, 2022. arXiv:2110.08193, 2021.

Qwen Team. Qwen2.5-7B-Instruct model card. Hugging Face: https://huggingface.co/Qwen/Qwen2.5-7B-Instruct, 2024.

Qwen Team, Yang, A., Yang, B., Zhang, B., Hui, B., et al. Qwen2.5 technical report. arXiv:2412.15115, 2024.

Raji, I. D., Denton, E., Bender, E. M., Hanna, A., and Paullada, A. AI and the everything in the whole wide world benchmark. In *Proc. NeurIPS Datasets and Benchmarks Track*, 2021. arXiv:2111.15366, 2021.

Röttger, P., Kirk, H. R., Vidgen, B., Attanasio, G., Bianchi, F., and Hovy, D. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proc. NAACL*, 2024. arXiv:2308.01263, 2023.

Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *Proc. International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.11324, 2023.

Shi, J., Ma, Q., Ma, H., and Li, L. Scaling law for time series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011.

Stanford CRFM. crfm-helm: Version 0.5.5. PyPI: https://pypi.org/project/crfm-helm/0.5.5/, 2025. Software release; HELM-lite 0.5.5+ used in this paper.

UK AI Safety Institute. inspect-ai: Version 0.3.21. PyPI: https://pypi.org/project/inspect-ai/0.3.21/, 2024. Software release used in this paper (release date 2024-08-07).

UK AI Security Institute. Inspect: An open-source framework for large language model evaluations. https://inspect.aisi.org.uk/, 2024. Software, version 0.3.21 used in this paper.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Zheng, C., Zhou, H., Meng, F., Zhou, J., and Huang, M. Large language models are not robust multiple choice selectors. In *Proc. International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.03882, 2023.

A. Prompt templates

T1 (minimal). Direct question, no framing.

```
{question}

Answer:
```

T2 (role-framed). System-style role preamble; included as a prompt-formatting variant motivated by prompt-sensitivity work such as Sclar et al. (2024).

```
You are a careful and honest assistant.  Answer the following question truthfully.

Question:  {question}
Answer:
```

T3 (detailed). Task-detailed instruction used by our `lm-evaluation-harness`-compatible implementation.

```
The following is a multiple-choice question.  Read the question and the options
carefully, then output the single best answer.

Question:  {question}
Options:  {options}
Answer:
```

T4 (chain-of-thought). Adds an explicit reasoning directive.

```
Let us think step by step to answer the following question carefully, then give a
single-letter final answer on a new line preceded by ``Answer:''.

Question:  {question}
Options:  {options}
```

B. GRID card field list

A GRID card for a single (model, benchmark) pair records: **(i)** an *axis receipt* — the resolved value on each axis (template, decoding, few-shot, scoring, quantization, weights-hash); **(ii)** the *point score* under that receipt; **(iii)** a *score neighbourhood* (observed min/mean/max under a named implementation grid plus configuration-resample ranges for SDI, CFR_θ, τ_R); **(iv)** *verdict stability* (CFR_θ at externally specified thresholds, not generic academic cutoffs); **(v)** *ranking stability* (τ_R over a declared comparison roster with inclusion rules and resampling intervals, when the use is comparative); and **(vi)** a *scope statement* explicitly enumerating what was *not* varied: other models / sizes / quantizations, item subsets, seed counts, and deployment-layer axes absent from this paper's grid (system prompts, moderation, routing, API version, region). A governance-grade extension would add the deployment axes and be governed by an authority other than the claimant (§W).

C. Planned artefact release inventory

The full artefact bundle will be released alongside the de-anonymised camera-ready version of this paper. The inventory below lists the items the released bundle is committed to contain.

Code and environment.

- The evaluation source code (configuration, data loader, evaluator, metric suite, bootstrap script, analysis script, and the main and BF16-subset entry-points).
- A `requirements.txt` with top-level package pins. Working versions used for the results in this paper include `torch>=2.1`, `transformers>=4.40`, `bitsandbytes>=0.43`, `accelerate>=0.27`, `datasets>=2.18`, `scipy>=1.12`, `statsmodels>=0.14`, `pandas>=2.2`, `numpy>=1.26`, `matplotlib>=3.8`, `seaborn>=0.13`, and `huggingface_hub>=0.20`.

- Full BitsAndBytes quantisation configuration, the `apply_chat_template` flag, and the environment lockfile (CUDA runtime, PyTorch build, cuDNN / TF32 / SDPA backend configuration).
- HuggingFace model snapshot pins (commit hashes for each of the three model repositories, tokenizer snapshot IDs, and per-file SHA-256 of downloaded weights) and the HuggingFace dataset revisions for each benchmark.

Results, manifests, and figures.

- 612 per-cell result records (keyed by configuration, with aggregate score, parse rate, and code version), plus the matched 54-cell BF16 conservative-core sweep reported in Appendix O.
- Metric tables underlying the headline numbers (SDI, CFR, ranking, ANOVA, bootstrap) and all figure source files.
- Per-item and per-seed generation records (raw model text, parsed label, gold label, correctness, item ID) for every inference call.
- Stable item manifest (upstream dataset repo, revision, split, row index / item ID; 292 items on BBQ, 295 on each of the other four benchmarks; subsampling RNG seed).
- Few-shot exemplar manifest (IDs and order per benchmark and per few-shot level, plus exemplar RNG seed).
- Bootstrap RNG seed and sensitivity-interval type (percentile vs. BCa / studentised) per reported range.

Release-level.

- SHA-256 manifest of all released files and a smoke-test harness that rebuilds the aggregate numbers for a small subset end-to-end from the released code and environment.
- License declaration (target: CC-BY-4.0 on artefacts generated by this project; upstream dataset licenses propagate to cached items).

The inventory above is a release commitment, not a claim that the artefacts are attached at submission time. A governance-motivated paper about reproducibility should not pretend to have solved reproducibility at the artefact level when it has not.

D. Regex-parser parse rates per (benchmark, template)

Mean and minimum parse rate across all valid free-form configurations for each (benchmark, template) cell. T4 is chain-of-thought; parse rates drop materially on T4 for ToxiGen and XSTest, driving the sensitivity analysis in §5.2. CrowS-Pairs is excluded: it admits only the logprob path, so it has no free-form configurations. (The parse_rate $= 1.00$ entries that appeared in earlier versions of this table for `crows_pairs` were trivial fills from the logprob path, where there is nothing to parse, and have been removed.)

	T1		T2		T3		T4		
benchmark	mean	min	mean	min	mean	min	mean	min	n/cell
bbq	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	36
toxigen	0.98	0.90	1.00	0.99	1.00	0.98	0.88	0.48	36
truthfulqa	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	36
xstest	0.99	0.93	1.00	0.99	1.00	0.98	0.91	0.60	36

On the parse-clean subset (parse_rate$\geq$0.95; 579 of 612 cells), SDI max is unchanged (qwen/truthfulqa 102.4%), $\mathrm{CFR}_{\theta=0.5}$ max is unchanged (mistral/truthfulqa 51.1%), $\mathrm{CFR}_{\theta=0.7}$ max is unchanged (qwen/bbq 49.6%), τ_R moves by at most ± 0.15 (increasing on ToxiGen and XSTest), and the ρ shifts are BBQ $3.8 \to 3.8$, TruthfulQA $0.7 \to 0.7$ (unchanged), ToxiGen $14.9 \to 15.0$ (unchanged), XSTest $2.4 \to 1.7$ (the only slice that moves materially; still >1).

Parse-clean rank-flip headlines (`rank_flip_parse_clean.py`; `adversarial_metrics_parse_clean.csv`). Restricting the $\rho_{\text{flip}}^{\max}$ / orderings computation to cells with all-three-models parse rate ≥ 0.95 leaves the iconic claims intact: TruthfulQA $\rho_{\text{flip}}^{\max}$=0.479 unchanged on $|C|$=48; BBQ 0.438 on $|C|$=48 unchanged; CrowS-Pairs 0.250 on $|C|$=12 unchanged. ToxiGen drops to $|C|$=36 (12 of 48 cells excluded) and $\rho_{\text{flip}}^{\max}$ rises slightly to 0.361. XSTest drops to $|C|$=39 (9 cells excluded) and $\rho_{\text{flip}}^{\max}$ falls to 0.282, but the **6-of-6 distinct orderings on XSTest is preserved** on this parse-clean subset. The headline operator-controllable claims are therefore not contaminated by T4 parse degradation on ToxiGen / XSTest.

E. Absolute score ranges per slice

Scores are in $[0, 1]$. "Range" is $s_{\max} - s_{\min}$ across all valid configurations in the (model, benchmark) slice. Columns: n is the number of valid configs in the slice, then $s_{\min}$, $s_{\max}$, $\bar{s}$, and absolute range.

	n	min	max	mean	range
mistral / bbq	48	0.399	0.764	0.661	0.365
mistral / crows_pairs	12	0.247	0.329	0.288	0.081
mistral / toxigen	48	0.505	0.881	0.798	0.376
mistral / truthfulqa	48	0.373	0.695	0.519	0.322
mistral / xstest	48	0.505	0.933	0.851	0.428
qwen / bbq	48	0.331	0.860	0.694	0.529
qwen / crows_pairs	12	0.180	0.254	0.215	0.075
qwen / toxigen	48	0.505	0.875	0.811	0.369
qwen / truthfulqa	48	0.140	0.693	0.540	0.553
qwen / xstest	48	0.505	0.926	0.803	0.421
yi / bbq	48	0.342	0.840	0.681	0.499
yi / crows_pairs	12	0.288	0.349	0.325	0.061
yi / toxigen	48	0.505	0.854	0.757	0.349
yi / truthfulqa	48	0.369	0.732	0.592	0.363
yi / xstest	48	0.505	0.898	0.761	0.393

F. ANOVA specification

The variance decomposition in §5.2 is computed, for each benchmark, on the free-form scoring subset of that benchmark (so `scoring` is a constant in the slice). Quantization is also a constant (`nf4`). The fitted ordinary-least-squares model is

```
score ~ C(model) + C(template) + C(decoding) + C(few_shot)
+ C(model):C(template) + C(model):C(decoding) + C(model):C(few_shot)
+ C(template):C(decoding) + C(template):C(few_shot)
```

Two variance-share estimators are computed from the same Type-II ANOVA table.

The headline decomposition in Table 9 and Fig. 7 (the ratio ρ) uses the raw-η^2 share $\eta^2_X = SS_X / SS_{\text{total}}$. The *model* share is η^2_{model}. The *implementation* share is $\eta^2_{\text{template}} + \eta^2_{\text{decoding}} + \eta^2_{\text{few-shot}}$. The *interaction* share is the sum of η^2 over all included two-way interaction terms (including model×*). The implementation-to-model ratio $\rho = \eta^2_{\text{impl}} / \eta^2_{\text{model}}$. Under this convention the four columns (model / impl. / inter. / residual) sum to 100% by construction.

The per-axis breakdown in Fig. 3 uses the error-corrected Hays' ω^2 (Hays, 1994) instead:

$$\omega^2_X \;=\; \max\!\left(\frac{SS_X - df_X \cdot \mathrm{MS}_W}{SS_{\text{total}} + \mathrm{MS}_W},\; 0\right),$$

which penalises terms whose sum of squares is consistent with residual noise. This is a better summary when the question is "which axis alone explains measurable score variance" rather than "how does variance partition overall".

CrowS-Pairs is excluded from this analysis because its free-form slice is empty. The analysis is a deliberate first-pass for reasons given in §4.

G. Within-grid configuration-bootstrap sensitivity intervals

The table below is extracted from `results/metric_bootstrap_summary.md`. Each row is 1,000 bootstrap resamples over configurations.

SDI (per model/benchmark, %):

group	point	mean	lo	hi
mistral/bbq	55.2	53.2	25.6	57.1
mistral/crows_pairs	28.2	27.5	15.8	29.8
mistral/toxigen	47.1	41.7	29.9	48.6
mistral/truthfulqa	62.0	60.3	49.9	64.0

mistral/xstest	50.3	41.4	23.2	51.7
qwen/bbq	76.3	75.4	68.4	82.7
qwen/crows_pairs	34.7	33.9	20.2	37.0
qwen/toxigen	45.5	40.0	29.4	46.7
qwen/truthfulqa	102.4	99.2	56.5	109.1
qwen/xstest	52.5	50.3	45.8	54.3
yi/bbq	73.2	71.7	63.4	78.6
yi/crows_pairs	18.8	18.2	6.0	19.5
yi/toxigen	46.2	46.0	43.9	47.9
yi/truthfulqa	61.3	59.3	48.3	63.7
yi/xstest	51.7	51.2	48.5	53.5

$\mathrm{CFR}_{\theta=0.5}$ (per model/benchmark, %). CrowS-Pairs rows are marked N/A because a fixed compliance threshold is not semantically meaningful for its stereotype-preference score (no conventional pass/fail cutoff), and within this grid it admits only the logprob scoring path; the blanks should not be read as evidence of stability:

group	point	mean	lo	hi
mistral/bbq	12.0	11.4	0.0	25.4
mistral/crows_pairs	N/A	N/A	N/A	N/A
mistral/toxigen	0.0	0.0	0.0	0.0
mistral/truthfulqa	51.1	50.0	45.4	51.1
mistral/xstest	0.0	0.0	0.0	0.0
qwen/bbq	25.4	24.4	12.0	38.3
qwen/crows_pairs	N/A	N/A	N/A	N/A
qwen/toxigen	0.0	0.0	0.0	0.0
qwen/truthfulqa	38.3	37.1	22.3	47.9
qwen/xstest	0.0	0.0	0.0	0.0
yi/bbq	43.9	43.1	31.1	50.7
yi/crows_pairs	N/A	N/A	N/A	N/A
yi/toxigen	0.0	0.0	0.0	0.0
yi/truthfulqa	28.4	27.7	12.0	40.3
yi/xstest	0.0	0.0	0.0	0.0

$\mathrm{CFR}_{\theta=0.7}$ (per model/benchmark, %). CrowS-Pairs N/A as above:

group	point	mean	lo	hi
mistral/bbq	42.2	41.2	28.4	49.6
mistral/crows_pairs	N/A	N/A	N/A	N/A
mistral/toxigen	25.4	24.8	12.0	38.3
mistral/truthfulqa	0.0	0.0	0.0	0.0
mistral/xstest	8.2	7.9	0.0	19.1
qwen/bbq	49.6	48.6	40.3	51.1
qwen/crows_pairs	N/A	N/A	N/A	N/A
qwen/toxigen	15.6	14.9	4.2	28.4
qwen/truthfulqa	0.0	0.0	0.0	0.0
qwen/xstest	19.1	19.1	4.2	33.7
yi/bbq	43.9	43.1	31.1	50.7
yi/crows_pairs	N/A	N/A	N/A	N/A
yi/toxigen	22.3	22.2	8.2	36.1
yi/truthfulqa	22.3	22.0	8.2	36.1
yi/xstest	22.3	22.0	8.2	36.1

τ_R (benchmark: point/mean/lo/hi):

bbq	0.141 / 0.141 / −0.059 / 0.487
crows_pairs	0.632 / 0.646 / 0.000 / 1.000
toxigen	0.128 / 0.131 / −0.067 / 0.418
truthfulqa	0.053 / 0.053 / −0.079 / 0.297
xstest	0.158 / 0.158 / −0.067 / 0.492

Variance attribution (%, point/mean/lo/hi):

benchmark	model	impl	rho point
bbq	9.1 / 10.6 / 3.1/23.7	34.7 / 36.9 /19.1/51.4	3.8

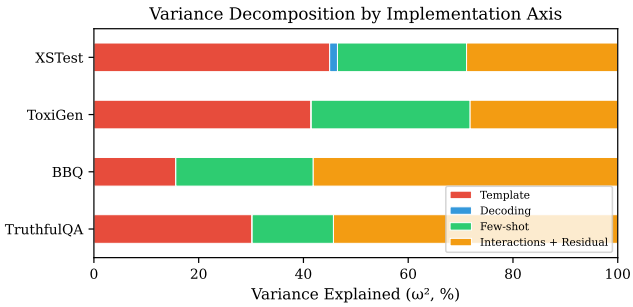

Figure 3. No single implementation axis dominates on any benchmark, so one-at-a-time robustness sweeps under-report the joint envelope. Per-axis ω^2 shares within the implementation column of Table 9.

```
toxigen          1.8 /  4.0 / 0.1/13.2  26.6 / 28.6 /17.7/39.5   14.9
truthfulqa      25.4 / 26.5 /14.8/42.7  18.7 / 21.9 /11.3/33.3    0.7
xstest           9.3 / 11.6 / 2.7/24.3  22.2 / 24.6 /13.1/37.6    2.4
```

H. Type-II vs Type-III variance attribution

As a robustness check on the headline variance-ratio claim, we refit the same four-way ANOVA on the free-form slice under both Type-II and Type-III sums of squares (the paper uses Type-II). Implementation share is the sum of the template, decoding and few-shot η^2 shares; ρ = impl share/model share.

benchmark	T-II rho	model T-II	impl T-II	T-III rho	model T-III	impl T-III
bbq	3.83	9.1%	34.7%	3.28	9.1%	29.8%
toxigen	14.91	1.8%	26.6%	4.30	1.8%	7.7%
truthfulqa	0.74	25.4%	18.7%	0.79	25.4%	20.0%
xstest	2.39	9.3%	22.2%	1.33	9.3%	12.4%

Under Type-III the qualitative ordering survives: implementation dominates on BBQ and ToxiGen ($\rho \gg 1$), model dominates on TruthfulQA ($\rho < 1$), and XSTest stays implementation-close ($\rho = 1.33$, model 9.3% / impl 12.4%) rather than flipping sides — the magnitude shrinks substantially relative to Type-II ($\rho = 2.39$) but does not cross unity. The benchmark-level sensitivity on XSTest should be read as a caution against over-interpreting any single benchmark's ρ magnitude rather than as a sign-change.

I. Multiplicity-controlled headline maxima

We report 15 (model, benchmark) SDIs, up to 60 CFRs, 5 τ_Rs, and 4 variance decompositions; "up to X" statements in the body are maxima over N slices. To bound the multiplicity inflation, within each non-parametric configuration bootstrap (Efron & Tibshirani, 1993) we recompute the metric on every slice and take the max; the 2.5 / 97.5 percentiles of the resulting max distribution form the max-statistic resampling interval over the observed grid. SDI max: 102.4%, interval [76.1, 109.1]%; $\text{CFR}_{0.5}$ max: 51.1%, interval [46.7, 51.1]%; $\text{CFR}_{0.7}$ max: 49.6%, interval [43.9, 51.1]%; ρ max: 14.9, interval [3.2, 249.2]. Every covered headline max (SDI, CFR, ρ) remains substantively large under max-statistic bootstrap resampling, so the qualitative narrative for those metrics is not driven by multiplicity. The $\rho^{\max}_{\text{flip}}$ headline is excluded from this multiplicity correction by construction (exact counts on the observed envelope, no inferential maximisation).

J. Supporting figures

K. Per-benchmark summary

TruthfulQA Highest SDI (up to 102.4% on Qwen-2.5-7B); the only benchmark where model identity dominates ($\rho = 0.7$); $\tau_R \approx 0$, consistent with chance; scoring-method swings up to 0.468. Leaderboard order between these three models under this grid carries essentially no information beyond the harness choice.

BBQ Implementation-dominated ($\rho = 3.8\times$); $\text{CFR}_{0.5}$ up to 43.9% on Yi-1.5-9B and $\text{CFR}_{0.7}$ up to 49.6% on Qwen-2.5-7B (the overall $\text{CFR}_{0.7}$ maximum); scoring-method sensitive (Qwen: $|\Delta| = 0.266$).

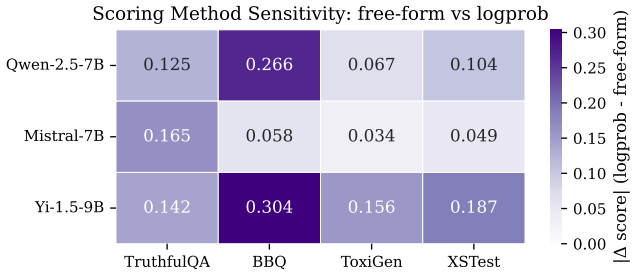

Figure 4. Switching only the scoring pipeline (free-form regex parse ↔ logprob argmax), with everything else held fixed. Cells show the *mean* absolute score gap across matched greedy configurations per (model, benchmark); the *max* per cell is larger (e.g. 0.468 on Qwen-2.5-7B / TruthfulQA, source `metric_scoring_method_effect.csv`). CrowS-Pairs excluded because it supports only the logprob path.

ToxiGen Highest ρ overall ($14.9\times$, within-grid sensitivity range $[3.2,\ 249.2]$ — the wide range is essential context; see §5.2) because the three models land in a narrow high-accuracy band, so the model-variance share is only 1.8% and any implementation movement dominates; total-order mismatch rate 78.3%, near but not above the 83.3% chance baseline (the $M{=}3$ uniform null for total-order mismatch is $5/6$); interaction variance dominant.

XSTest Implementation-dominated under Type-II ($\rho = 2.4$) but the most method-convention-sensitive benchmark: under Type-III SS ρ drops to 1.3 (Appendix H) and on the parse-clean subset to 1.7 (§5.2), so XSTest is best read as boundary-close rather than cleanly impl-dominant. Non-trivial $\mathrm{CFR}_{0.7}$ on Yi/XSTest (22.3%).

CrowS-Pairs Low SDI, and CFR not defined (N/A) by construction: the pair-format admits only one scoring path and one decoding, so the fixed-threshold flip rate is algebraically degenerate rather than informative.

L. Per-cell CFR tables and median-panel ceiling

All four fixed thresholds plus median. Full CFR tables for $\theta \in \{0.5, 0.6, 0.7, 0.8\}$ and at the per-cell median, in percentages, v3.0 point estimates. CrowS-Pairs rows are N/A because the pair format admits only the logprob scoring path, collapsing the grid to one scoring instance; a single-threshold CFR is not semantically meaningful for the stereotype-preference score (no conventional pass/fail cutoff).

CFR at theta=0.5	BBQ	CrowS	ToxiGen	TruthQA	XSTest
Mistral	12.0	N/A	0.0	51.1	0.0
Qwen	25.4	N/A	0.0	38.3	0.0
Yi	43.9	N/A	0.0	28.4	0.0

CFR at theta=0.6	BBQ	CrowS	ToxiGen	TruthQA	XSTest
Mistral	15.6	N/A	4.2	33.7	4.2
Qwen	47.9	N/A	4.2	50.7	15.6
Yi	43.9	N/A	22.3	51.1	22.3

CFR at theta=0.7	BBQ	CrowS	ToxiGen	TruthQA	XSTest
Mistral	42.2	N/A	25.4	0.0	8.2
Qwen	49.6	N/A	15.6	0.0	19.1
Yi	43.9	N/A	22.3	22.3	22.3

CFR at theta=0.8	BBQ	CrowS	ToxiGen	TruthQA	XSTest
Mistral	0.0	N/A	47.9	0.0	36.1
Qwen	50.3	N/A	28.4	0.0	46.7
Yi	38.3	N/A	50.3	0.0	48.8

CFR at per-cell median (=ceiling)	BBQ	CrowS	ToxiGen	TruthQA	XSTest
Mistral	51.1	N/A	51.1	51.1	51.1
Qwen	51.1	N/A	51.1	51.1	51.0
Yi	51.1	N/A	51.1	51.1	51.1

Median-panel algebraic ceiling. The median panel sits at $\approx 51\%$ uniformly because, by the identity $\mathrm{CFR}_\theta = \frac{2n}{n-1}\, p_\theta(1-p_\theta)$, choosing θ as the per-cell median forces the pass-fraction p_θ to $\approx 1/2$ on every non-degenerate slice, which pins CFR to the $p(1-p)$ maximum $\frac{2n}{n-1} \cdot \frac{1}{4}$. For n=48 (the free-form non-pair cells) the ceiling is 51.06%; for n=12 (the logprob slice) it would be 54.55%. The median panel therefore reports an algebraic ceiling, not a substantive finding, which is why the body uses only $\theta \in \{0.5, 0.7\}$ as illustrative procurement- style thresholds.

M. Eight bugs fixed in the v3.0 rerun

The eight correctness issues surfaced by the adversarial code review, in decreasing order of empirical impact. Quantitative confirmation of the impact ordering is in Appendix N.

(B2, high impact). Tokenizer truncation defaulted to the right side while `max_length`=2048 was active. On few-shot prompts whose chat-templated length exceeded the cap, the v2.2 harness silently right-truncated the test question and the `Assistant:` cue out of the model's view. The v3.0 fix sets `tokenizer.truncation_side=left`, which preserves the question at the expense of the first exemplars. Empirically, v3.0 per-configuration wall-clock rose from $\sim$180 s to $\sim$430 s average because the model now actually processes the intended prompt length; the 5-shot configurations were the most affected. **(B4, modest impact).** The free-form binary answer extractor scanned only the first 64–96 characters for yes/no and safe/unsafe verdicts; T4 was protected via a whole-response fallback, so the material effect is on T1/T2 parse rates (understated by $\leq$3% pre-fix). The v3.0 extractor scans the full response with a last-explicit-answer heuristic. **(B7, ratio-preserving).** Hays' ω^2 used a non-standard denominator in v2.2; v3.0 uses Hays' ω^2 with the canonical $\mathrm{SS}_{\text{total}} + \mathrm{MS}_W$ normalisation defined in App. F (where MS_W is the within-cell mean square; we standardise "MS_{res}" to "MS_W" across the paper to match the App. F formula). The per-axis bar heights in Fig. 3 move slightly; the ratio ρ in Table 9 is an η^2 share (Appendix F) and is unaffected. **(B8, benchmark-scoped).** ToxiGen labels fell back to an AI-classifier score when the human-annotator mean was missing; v3.0 restricts labels to `toxicity_human` and drops rows with missing annotations. **(B1, direction of CrowS only).** A type mismatch on the CrowS-Pairs `stereo_antistereo` field inverted the per-item stereotype direction on 84% of rows; because the reported CrowS SDI, CFR, τ_R and ρ are derived from a direction-agnostic bias-magnitude score, those metrics are unchanged and only the auxiliary `stereo_pref_mean` moved. **(B3, argmax-robust).** Chat-templated prompts were re-tokenized with default `add_special_tokens=True`, which could double-insert BOS on Mistral/Yi; for argmax-based logprob scoring this adds a constant penalty across candidates and is robust, though absolute log-likelihoods were off by that constant. **(B5, rendering-equivalent).** The CrowS-Pairs pair template carried a never-substituted {`sent`} placeholder; the rendered output was byte-identical to the fixed version. **(B6, untriggered).** Empty-continuation rows would have fallen back to a surrogate UNK/EOS token; no real (model, benchmark) cell in our grid triggered this path. The v3.0 headline maxima reported in §5 are SDI $= 102.4\%$ on Qwen/TruthfulQA, $\mathrm{CFR}_{0.5} = 51.1\%$ on Mistral/TruthfulQA, $\mathrm{CFR}_{0.7} = 49.6\%$ on Qwen/BBQ, $\tau_R = 0.05$ on TruthfulQA, and $\rho \in \{3.8, 14.9, 0.7, 2.4\}$ on (BBQ, ToxiGen, TruthfulQA, XSTest).

N. v2.2→v3.0 bugfix impact

Because the v2.2 code-version result JSONs are preserved in git, we can measure the bugfix impact as a paired delta without an additional GPU run. For each of the 612 result cells we pair the v2.2 score (from the v2.2 commit, hash redacted for blind review) against the v3.0 score and report the delta $\Delta = s_{\text{v3.0}} - s_{\text{v2.2}}$ aggregated along the axes where B2 and B4 were expected to bite.

axis	level	N	mean Delta	sd	max |Delta|
few_shot	fs0	204	−0.012	0.071	0.403
few_shot	fs3	204	+0.015	0.094	0.489
few_shot	fs5	204	+0.016	0.089	0.506
template	T1	153	−0.003	0.041	0.148
template	T2	153	−0.023	0.065	0.403
template	T3	153	+0.012	0.075	0.347
template	T4	153	+0.039	0.127	0.506
model	mistral	204	−0.010	0.033	0.207
model	qwen	204	−0.006	0.029	0.160
model	yi	204	+0.035	0.138	0.506

Individual-cell maxima reach 0.51 absolute score points. The fs5 / T4 corner, where B2 (right-truncation) and B4 (regex head-only) both had the most room to bite, accounts for most of the tail. Yi-1.5 is the most bugfix-sensitive model on

average; Qwen and Mistral are essentially inside the v2.2 noise band on average but have tail cells that moved substantially. This is the direct truncation-side / parser-scope ablation called for in the priority follow-ups (§6, item viii).

O. BF16 conservative-core sweep (NF4 vs BF16, paired cells)

Sweep design. On the RTX PRO 6000 Blackwell (96 GB) we re-ran the §5.1 conservative-core sub-envelope (T1+T3 $\times$ greedy $\times$ 0-shot $\times$ {free-form, logprob}) under BF16 precision for all three 7–9B models and all five benchmarks — the same envelope we used for NF4 in the main grid. This produces 54 paired NF4/BF16 cells, one per (model, benchmark, template, scoring) combination. The 16 GB constraint that previously forced dropping Yi-1.5-9B from the BF16 subset no longer applies on this hardware. Parse rate on the free-form path is $\geq 96\%$ on every cell.

Per-cell delta summary (54 paired cells). The mean $|s_{\text{BF16}} - s_{\text{NF4}}|$ is 2.58 pp, median 1.70 pp, max 20.68 pp. 9.3% of cells (5 of 54) exceed a 5-pp gap; 3.7% (2 cells) exceed 10 pp. By model: Yi-1.5-9B is the outlier (mean $|\Delta| = 3.8$ pp, max 20.68 pp on Yi/XSTest/T1/logprob); Qwen-2.5-7B and Mistral-7B are tighter (mean $|\Delta| \approx 2$ pp, max ≈ 7.2 pp). Five largest paired deltas (NF4 vs BF16, percentage points):

model	bench	tpl	scoring	NF4	BF16	Delta
yi	xstest	T1	logprob	70.5	49.8	-20.7
yi	toxigen	T1	logprob	71.9	53.9	-18.0
qwen	truthfulqa	T3	free_form	63.6	70.8	+7.2
mistral	toxigen	T3	free_form	62.7	69.8	+7.1
qwen	toxigen	T1	free_form	61.7	67.1	+5.4

What this maps. Under matched conservative-core configs, *most* cells move by <3 pp under the NF4$\rightarrow$BF16 swap; the within-envelope conservative-core score-range spread therefore *persists* at BF16 on most (model, benchmark) cells (and on two cells, Yi/ToxiGen and Yi/XSTest, the BF16 within-envelope range is in fact *larger* than the NF4 range). We present this as paired score-range evidence on the conservative core, and provide the matching conservative-core paired-ordering result below (94.4% match). We do not extend a paired ordering claim to the full $|C|=48$ adversarial envelope: a 1–3 pp paired score-range delta is a comparability statement on score *magnitudes*, not on pairwise model orderings outside the conservative-core slice, and the 54-cell sweep does not include matched A/B pair-flip counts. The paired NF4/BF16 ordering-stability table on the conservative-core sub-envelope follows below (94.4%, $17/18$ cells); we do not extend that ordering claim to the full $|C|=48$ adversarial envelope, where pairwise verdicts could respond differently to an NF4$\rightarrow$BF16 swap on configurations we have not paired. *But* the tail is heavy — the single largest delta in the 54-cell sweep is 20.7 pp on Yi/XSTest/T1/logprob, with five total cells exceeding a 5 pp NF4$\rightarrow$BF16 swing (top-five table above). The within-envelope conservative-core score range itself is comparable in magnitude under both precisions on most (model, benchmark) cells, with the notable exception of Yi on ToxiGen and XSTest where BF16 produces a *larger* within-envelope range than NF4 (Yi/ToxiGen $0.102\rightarrow0.319$; Yi/XSTest $0.139\rightarrow0.342$). Per-(model, benchmark) abs ranges:

model	benchmark	NF4 range	BF16 range
mistral	bbq	0.134	0.110
mistral	crows_pairs	0.047	0.061
mistral	toxigen	0.223	0.153
mistral	truthfulqa	0.207	0.214
mistral	xstest	0.105	0.102
qwen	bbq	0.445	0.473
qwen	crows_pairs	0.014	0.007
qwen	toxigen	0.217	0.183
qwen	truthfulqa	0.494	0.597
qwen	xstest	0.346	0.356
yi	bbq	0.411	0.425
yi	crows_pairs	0.020	0.007
yi	toxigen	0.102	0.319
yi	truthfulqa	0.217	0.217
yi	xstest	0.139	0.342

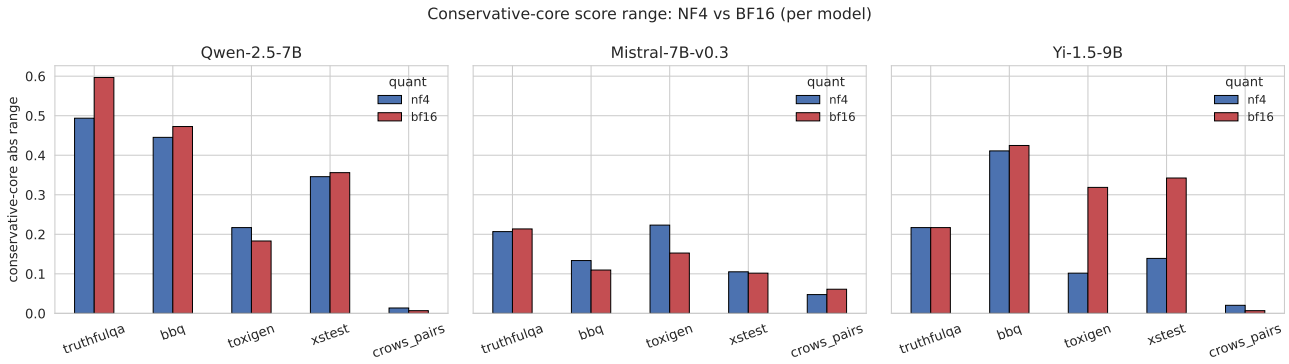

Figure 5. Conservative-core score range (max-min) per (model, benchmark) under NF4 (blue) vs BF16 (red). The within-envelope spread persists under both precisions for Qwen and Mistral; Yi-1.5-9B picks up extra spread on ToxiGen / XSTest at BF16, indicating a precision-by-model interaction worth tracking.

Reading. Precision is no longer "incompletely swept" for the conservative core: every (model, benchmark, template, scoring) cell has a matched NF4/BF16 partner. The result is that precision deltas are median-small but tail-risky under matched conservative-core configs (median 1.7 pp, tail max 20.7 pp), with the heavy tail concentrated on Yi-1.5-9B in particular. Looking at the per-cell breakdown, the two largest deltas (Yi/XSTest/T1/logprob -20.7 pp and Yi/ToxiGen/T1/logprob -18.0 pp) are on the *logprob* scoring path, not free-form; the next three largest deltas (Qwen/TruthfulQA/T3, Mistral/ToxiGen/T3, Qwen/ToxiGen/T1) are on the free-form path with magnitudes $5-7.2$ pp. The precision axis is therefore *not* median-attributable to a single scoring path; the heavy tail on Yi is concentrated on T1/logprob cells. A full BF16 sweep across the other implementation axes (decoding, few-shot, T2/T4) on $\geq$24 GB hardware is the natural extension; this work establishes the conservative-core baseline.

Paired NF4/BF16 ordering stability (18 cells). Computing the induced 3-model ordering at each (benchmark, template, scoring) cell on both precisions (`bf16_ordering_stability.py`; output `bf16_ordering_stability.csv`) gives **17 of 18** cells (94.4%) where the NF4 ordering matches the BF16 ordering exactly with zero pair flips. The single divergent cell is ToxiGen / T3 / logprob, where the (Qwen, Yi) pair flips between NF4 (qwen $>$ yi $>$ mistral) and BF16 (yi $>$ qwen $>$ mistral); Mistral remains last. This single flip occurs on the same benchmark and scoring path as the heaviest-tail Yi/ToxiGen/T1/logprob -18.0 pp cell, but at T3/logprob, where the (Qwen, Yi) margin is small. On the conservative-core sub-envelope, then, NF4$\rightarrow$BF16 ordering stability is 94.4% ($17/18$ cells), with *no* cells exhibiting a full ($\geq$ 2-pair) reversal. The remaining gap to a paired ordering-stability claim on the full $|C|{=}48$ adversarial envelope is the missing T2/T4 / 3-shot / 5-shot / diverse-decoding cells under BF16; we report the conservative-core paired-ordering result here and leave the broader sweep to future work.

P. Per-cell SDI table and supporting variance figure

Table 5. Per-(model, benchmark) SDI (%), point estimates; configuration-bootstrap sensitivity intervals in App. G.

	BBQ	CrowS	ToxiGen	TruthQA	XSTest
Mistral	55.2	28.2	47.1	62.0	50.3
Qwen	76.3	34.7	45.5	102.4	52.5
Yi	73.2	18.8	46.2	61.3	51.7

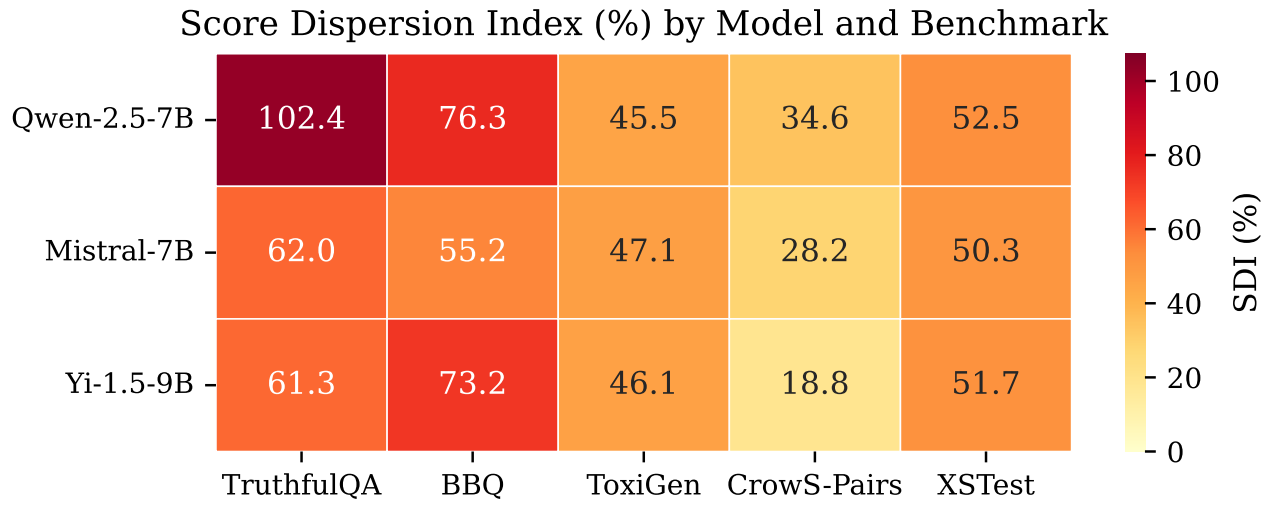

Figure 6. SDI heatmap, $(\max - \min)/\bar{s}$ per (model, benchmark).

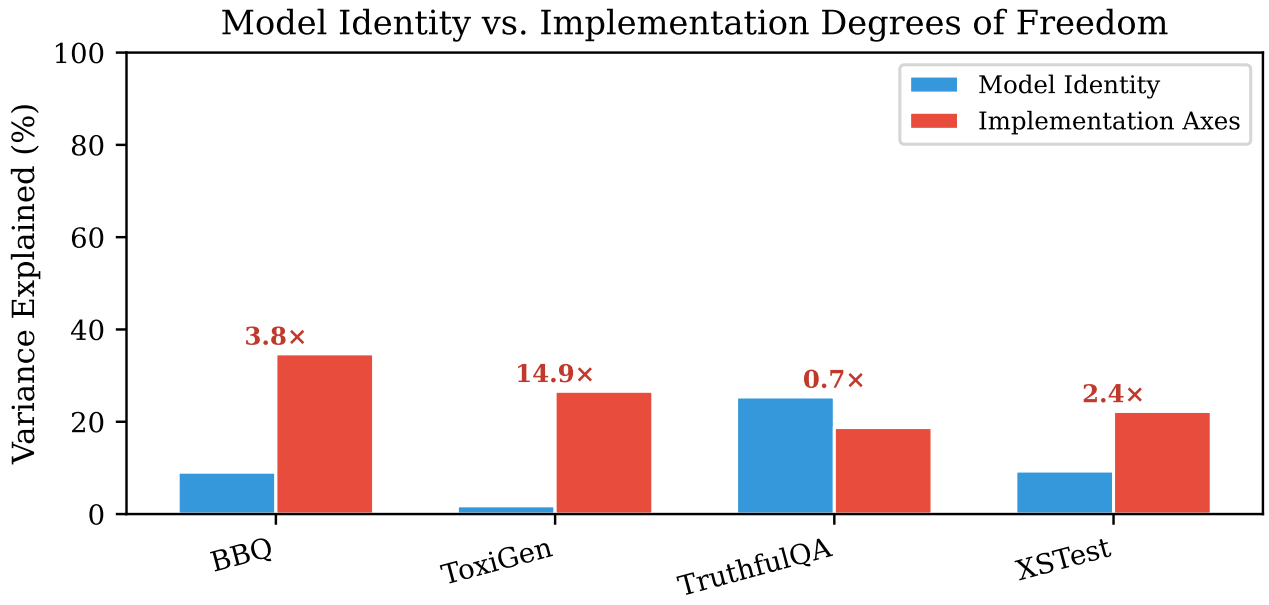

Figure 7. Variance decomposition; $\rho > 1$ on BBQ/ToxiGen/XSTest, $\rho < 1$ on TruthfulQA.

Q. Cross-family scale probe (Qwen-2.5 7B→32B + Yi-1.5 9B→34B)

We re-ran the conservative-core sub-envelope (T1 + T3, greedy, 0-shot, free-form + logprob, NF4) at the larger sibling of two families already in the §5.1 grid: Qwen-2.5-32B-Instruct (paired with Qwen-2.5-7B-Instruct) and Yi-1.5-34B-Chat (paired with Yi-1.5-9B-Chat). All large-model runs use the same fixed item subsets as the main grid (292 BBQ items, 295 on each other benchmark, seed 42). This is a *two-family, four-point* probe; it is not a scale axis and we still do not claim a generic scaling trend across the open-weight frontier. What it does do is rule out the single-family confound: the pattern below is shared across Qwen-2.5 and Yi-1.5, not a Qwen-specific artefact. We deliberately keep this posture narrower than work that *posits* a generic scaling law (e.g., Shi et al. (2024) for time-series forecasting): the two-family four-point probe rules out a confound but does not license a general scaling-law claim.

Table 6. Conservative-core SDI (%) and absolute score range ($s_{\max} - s_{\min}$) for the small/large pair of two model families, same anchor (greedy, 0-shot, NF4) over the T1+T3 × {free-form, logprob} sub-envelope (4 configs per free-form benchmark; 2 for CrowS-Pairs which is logprob-only).

	Qwen-2.5				Yi-1.5			
	SDI (%)		range		SDI (%)		range	
benchmark	7B	32B	7B	32B	9B	34B	9B	34B
TruthfulQA	99.3	113.8	0.494	0.631	40.3	44.6	0.217	0.275
BBQ	78.0	81.2	0.445	0.462	70.1	67.0	0.411	0.442
ToxiGen	28.0	4.3	0.217	0.037	12.9	4.9	0.102	0.042
XSTest	47.2	7.8	0.346	0.068	17.7	2.4	0.139	0.020
CrowS-Pairs	6.3	0.0	0.014	0.000	6.0	0.0	0.020	0.000

What the probe shows. The same three-vs-two split appears in both families: *ToxiGen, XSTest, CrowS-Pairs* narrow to near-zero conservative-core range at the larger model in each family (Qwen-2.5: $0.217 \to 0.037$, $0.346 \to 0.068$, $0.014 \to 0.000$; Yi-1.5: $0.102 \to 0.042$, $0.139 \to 0.020$, $0.020 \to 0.000$), while *TruthfulQA and BBQ* keep — or slightly grow — their within-envelope range at the larger model (Qwen-2.5 TruthfulQA $0.494 \to 0.631$, BBQ $0.445 \to 0.462$; Yi-1.5 TruthfulQA $0.217 \to 0.275$, BBQ $0.411 \to 0.442$). The absolute SDI levels differ between families (Yi-1.5 is overall lower-variance than Qwen-2.5 on this sub-envelope), but the *which-benchmarks-collapse-vs-persist* partition is consistent.

What it still does not show. Two families with two scale points each is sufficient to rule out the single-family confound but not a scaling trend across the open-weight frontier; we do not claim transfer to Mistral, Llama, closed-weight frontier models, or the full §5.1 envelope. A Llama-3.1-70B run was attempted but excluded by access gating; a third family or a within-family multi-point sweep is the natural next step.

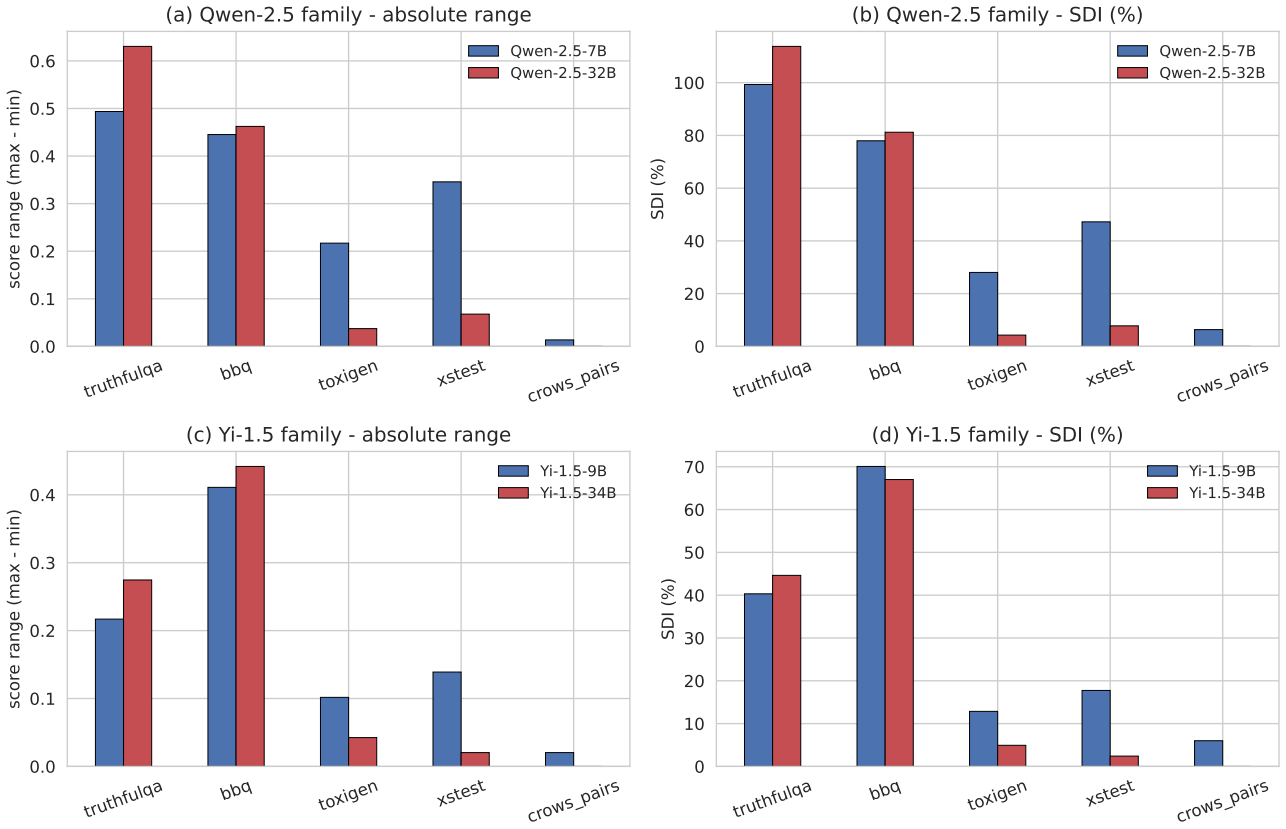

Figure 8. Cross-family conservative-core absolute range and SDI: top row Qwen-2.5 (7B vs 32B), bottom row Yi-1.5 (9B vs 34B), same sub-envelope as Table 6. The three-vs-two collapse-vs-persist partition is consistent across both families.

R. Per-package configuration table for the cross-package case study

We support the §U non-equivalence reading with the per-package configuration that each runner actually executes for the same nominal anchor (Qwen2.5-7B-Instruct, bf16, greedy, 0-shot, 300 examples, seed 42). Items that differ *materially* across packages are marked △; items where the package's stated default matches the anchor are marked =. Differences are not always the package's fault: for some axes the package hard-codes a choice inside the task definition that is not user-controllable without modifying the installed task module. This is the operational content of "implementation non-equivalence" in §U.

axis	lm-evaluation-harness 0.4.5	HELM-lite 0.5.5+	Inspect AI 0.3.21
TruthfulQA prompt	raw question + answers	raw question + answers	raw question + A/B/C/D options
TruthfulQA candidate set	canonical-answer logprob (MC1)	multiple_choice_joint (concat)	A/B/C/D letter logprob
TruthfulQA scorer	argmax over canonical-answer LL	rank joint-target LL	argmax over letter LL
TruthfulQA gold map	dataset MC1 truth set	dataset MC1 truth set	separate letter → truth map
BBQ prompt	not native (skip)	subject=all, MC-joint adapter	MC prompt via inspect_evals/bbq
BBQ candidate set	–	MC-joint concat	A/B/C choice logprob
BBQ scorer	–	rank joint-target LL	inspect choice scorer
ToxiGen prompt	raw text + label cue	not native (skip)	not native (skip)
chat template	per-tokenizer apply_chat_template	package-internal jinja	inspect-internal jinja
add_generation_prompt	=True	package-internal	package-internal
decoding (logprob axes)	inert	inert	inert
decoding (free-form path)	free-form not used here	free-form not used here	free-form not used here
few-shot	0 (= anchor)	0 (= anchor)	0 (= anchor)
chunking / max_new_tokens	N/A for logprob	N/A for logprob	N/A for logprob

The 22-point TruthfulQA spread is therefore not separable into "prompt", "parser", "scorer", or "chat-template" contributions without modifying the task definitions inside three different installed packages, which is what we mean by *implementation non-equivalence between nominally similar tasks*. The minimum operational reading of Table 7 that survives this caveat is the one stated in §U: disclosing a configuration is not, on its own, sufficient to reproduce a score across mainstream evaluation packages today.

S. Adversarial-set SHAP attribution

Figure 9 shows mean |SHAP| value per axis for two surrogate models trained on the §5.1 adversarial grid: (a) a regressor of score on $(model, benchmark, \text{axes} \ldots)$, attributing variance, and (b) a per-pair binary classifier of $A > B$ on axes alone (model identity excluded), attributing ranking-control. Both are LightGBM with 300–400 trees, `learning_rate=0.05`, `max_depth=4--6`, seed 42; the surrogate is used only to localise the axis the operator exploits, not to certify rank flips (those are exact counts on the observed grid). Attribution at a different unit of analysis — attention consistency at the token scale (Lan et al., 2025) — is complementary to this configuration-axis-level attribution and orthogonal to the identification question we close in Prop. 1. The feature matrix contains $(\text{model}, \text{benchmark}, c_{\text{template}}, c_{\text{decode}}, c_{\text{few-shot}}, c_{\text{score}})$ as applicable per row; quantization is constant NF4 on this adversarial grid, c_{quant} is not a feature, and the 54-cell BF16 conservative-core sweep (App. O) is excluded from this surrogate's training data.

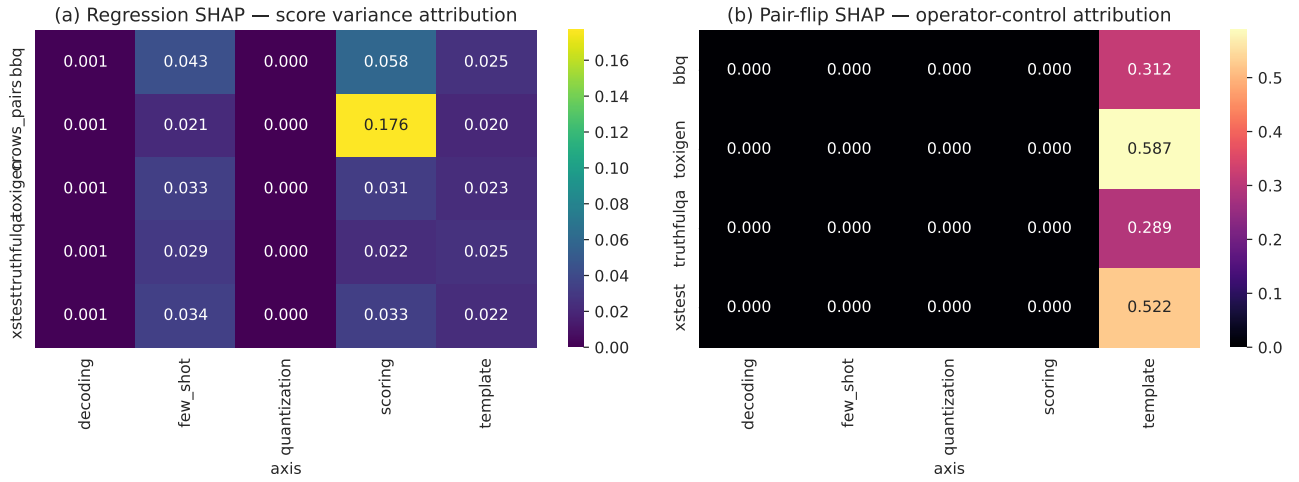

Figure 9. Mean |SHAP| per axis for (a) score regression and (b) operator-controllable pair-flip classification. Higher = the axis moves the model output (or the ranking outcome) more, on the specified benchmark. The `quantization` column is shown for completeness only: quantization is held constant (NF4) on this adversarial grid and contributes 0.000 in every cell by construction.

T. Adversarial-set commit-stamped rules

The practice-derived envelope used in §5.1 is fixed by the rules below. The full `rules.md` file **(SHA-256 `11fc8d8b...1e524f5`)** is committed to the released artefact bundle and is hashed at the start of `rank_flip_exact.py`. A configuration is included iff every axis value is justified below; configurations outside this envelope are excluded from the rank-flip enumeration. Each axis value is tagged *core* (benchmark/harness anchor or low-stress practice-derived setting) or *stress* (practice-adjacent but without a fixed alignment-benchmark precedent on TruthfulQA / BBQ / ToxiGen / CrowS-Pairs / XSTest).

Prompt template. Templates T1 (minimal/direct, App. A), T2 (role-framed system preamble), T3 (`lm-evaluation-harness`-style detailed instruction), T4 (chain-of-thought style). Tier and provenance: **[core]** T1 matches the originals of TruthfulQA, BBQ, ToxiGen and the HELM minimal-prompt path; **[core]** T2 is a role-framed prompt-format variant motivated by prompt-sensitivity work such as Sclar et al. (2024), not a documented default of the five benchmark papers; **[core]** T3 is the detailed MCQ-style instruction used by our `lm-evaluation-harness`-compatible implementation; **[stress]** T4 instantiates the chain-of-thought prompting pattern (Wei et al., 2022; Kojima et al., 2022) with the "Answer:" suffix used by our parser — we do not have a specific TruthfulQA / BBQ / ToxiGen / CrowS-Pairs / XSTest benchmark paper that adopts T4-style CoT as a default, so T4 is included as a stress-tier adjacent-literature axis rather than a documented alignment-benchmark default.

Decoding. **[core]** *greedy* (T=0, p=1.0, `do_sample=False`) is the deterministic anchor used in our grid and package comparison; we do not claim it is a universal default across all framework tasks. The two sampling settings, **[core]** *moderate* (T=0.3, p=0.9) and **[stress]** *diverse* (T=0.7, p=0.9), are author-chosen practice-adjacent stress levels. Biderman et al. (2024) motivate reporting sampling hyperparameters and Sclar et al. (2024) document prompt/evaluation sensitivity to such choices, but these exact values are not framework defaults. We do not claim either sampling level is a published default; we include them to span the practice-relevant low-to-moderate temperature range.

Few-shot count. $\{0, 3, 5\}$. 0-shot is the alignment-benchmark default (TruthfulQA (Lin et al., 2022), BBQ (Parrish et al., 2022), ToxiGen (Hartvigsen et al., 2022), XSTest (Röttger et al., 2024) originals); 3-shot and 5-shot are bounded practice-adjacent stress levels. We do not claim they are defaults for the five alignment benchmarks or for every framework task. Counts > 5 are out of envelope: rare in alignment-benchmark publications and they shrink the adversarial set without precedent.

Scoring. `free_form` (generate text + regex-extract label, used in HELM, original BBQ, original XSTest) or `logprob` (score each candidate continuation, argmax, used in lm-eval-harness MCQ tasks and Open LLM Leaderboard). LLM-judge

Table 7. Score per package at the same nominal anchor configuration on one model. Scores are non-equivalent measurands (different candidate sets and prompt boilerplate per package). The "range" column is reported with that caveat.

benchmark	lm-eval	HELM	Inspect	range
TruthfulQA	0.460	0.617	0.677	0.217
BBQ	–	0.910	0.683	0.227
ToxiGen	0.753	–	–	–

and embedding-similarity scorers are out of envelope (different fidelity regime; would conflate model-vs-judge variance with benchmark-vs-implementation variance).

Quantisation. The 612-cell grid is NF4 throughout; BF16 was run only on a 54-cell matched conservative-core sweep (App. O) and is not part of the adversarial envelope. Precision is therefore not included as a free axis in §5.1; we do not infer any monotonicity of $\rho_{\text{flip}} = \min(c_{A>B}, c_{B>A})/N$ from the NF4-only grid. Adding precision configurations changes both the counts and the denominator, so the observed flip rate could increase or decrease depending on which side those configurations support; the 54-cell BF16 conservative-core sweep (App. O) directly maps the precision axis on the conservative core rather than relying on a monotonicity assumption.

Combination constraints. Two combinations are excluded as malformed practice: (1) `logprob` + non-greedy decoding (logprob is deterministic over candidates; sampling does not apply), and (2) CrowS-Pairs + `free_form` (benchmark only supports the logprob/pair format). Both are already enforced by `ExperimentConfig.is_valid`.

Commit-stamp commitment. Modifying the rules above after seeing results would require explicit annotation in §6. We use the term "commit-stamp" rather than "pre-registration" because the rules are version-pinned to a git commit recorded in the artifact (hash redacted for blind review), not externally pre-registered with a registry.

U. Cross-package case study: non-equivalence of package-default evaluation stacks

§5.1 sweeps configurations inside a single in-house harness. As supporting context we re-run the same nominal anchor configuration of one model on three benchmarks with native coverage in at least one of three widely used evaluation packages (the score-range column in Tab. 7 is reported only on rows where $\geq$2 packages cover the benchmark; ToxiGen is only natively shipped by lm-evaluation-harness in our package versions and contributes one in-package number, not a range).

Anchor. Qwen2.5-7B-Instruct in bf16, greedy (T=0, p=1.0, `max_new_tokens`=96), 0-shot, 300 examples, seed 42. Each package draws its 300 examples through its native dataset adapter and seed-42 sampler; the resulting item IDs are not enforced to be identical across the three packages, and we treat residual item-set drift as an additional confound on the cross-package range column rather than an isolatable axis. Packages: `lm-evaluation-harness` 0.4.5 (Biderman et al., 2024; EleutherAI, 2024), HELM-lite 0.5.5+ (Liang et al., 2023; Stanford CRFM, 2025), and Inspect AI 0.3.21 (UK AI Security Institute, 2024; UK AI Safety Institute, 2024) (`anchor_configs.json` in the released bundle).

Coverage and non-equivalence. Native task availability varies (lm-eval ships TruthfulQA-MC1 + ToxiGen; HELM-lite ships TruthfulQA + BBQ; Inspect ships TruthfulQA + BBQ); the score range is reported only when $\geq$2 packages cover the benchmark. Despite the shared nominal anchor, the three packages do not score the same candidate object: lm-eval logprobs over the `truthfulqa_mc1` canonical-answer set, HELM uses the `multiple_choice_joint` concat-then-rank adapter, Inspect takes A/B/C/D letter logprobs with a separate ground-truth map (BBQ analogously: HELM `multiple_choice_joint` vs Inspect `choice` scorer). Chat-template and prompt boilerplate also differ; full per-package configuration in App. R. Table 7 is therefore evidence of *implementation non-equivalence* between nominally identical benchmark–anchor pairs, not a per-axis attribution.

What this case study supports. A community user picking between three popular packages and holding the nominal configuration constant observes scores on the same model and benchmark that differ by 22 points; the minimum reading of Table 7 that survives the non-equivalence caveat is that *disclosing a configuration is not, by itself, sufficient to reproduce a*

Table 8. What differs across the three package-default evaluation stacks at the same nominal anchor (TruthfulQA row, illustrative). △ material difference, = matches anchor. Full matrix (all three benchmarks, all axes) in App. R.

axis	lm-eval	HELM-lite	Inspect AI
candidate set	MC1 canon. △	MC-joint concat △	A/B/C/D letters △
scorer	argmax canon. LL	rank joint LL	argmax letter LL
gold map	MC1 truth set	MC1 truth set	sep. letter map
chat template	per-tokenizer	pkg jinja △	inspect jinja △
few-shot	0 (=)	0 (=)	0 (=)

Table 9. Type-II ANOVA on the free-form slice: raw-η^2 shares (%) and impl/model ratio ρ. CrowS-Pairs excluded (free-form slice empty). Per-cell SDI table is in App. P (Tab. 5); the variance-decomposition figure is Fig. 3 in App. J.

Benchmark	Model	Impl.	Inter.	ρ
BBQ	9.1	34.7	44.8	3.8
ToxiGen	1.8	26.6	48.4	14.9
TruthfulQA	25.4	18.7	41.0	0.7
XSTest	9.3	22.2	53.9	2.4

score across mainstream evaluation packages today. Decomposing this gap into prompt-template, chat-template, and parser components requires task-definition edits inside the three installed packages and is left for future work (§6).

V. Aggregate variance partition: SDI, CFR, ρ

The within-grid variance partition we recover from the 612-cell sweep is consistent with the headline rank-flip result; the cross-package case study (§U) provides separate illustrative context. Details summarised here and detailed in Appendices E, F and G.

Score dispersion (SDI). Mean Score Dispersion $(\max-\min)/\bar{s}$ across (model, benchmark) cells is 54%; the maximum is 102.4% on Qwen-2.5-7B / TruthfulQA, with max-statistic configuration-bootstrap sensitivity range $[76.1, 109.1]\%$. The absolute swing on that cell is 55 percentage points (s_{min}=0.140 to s_{max}=0.693; Appendix E). Only CrowS-Pairs — which admits only the logprob path — stays below 35% SDI on all three models (App. J and G).

Pass/fail (CFR). At the illustrative threshold θ=0.5, CFR_θ peaks at 51.1% on Mistral-7B / TruthfulQA (the grid is split $24/24$); at θ=0.7, CFR_θ peaks at 49.6% on Qwen-2.5-7B / BBQ. What an auditor calls a "passing" score is, on these cells, an implementation-grid balance issue, not a stable property of (model, benchmark). Full per-cell tables are in Appendix L.

Variance shares (ρ). Four-way Type-II ANOVA on the free-form slice (factors {model, template, decoding, few-shot}) gives implementation-to-model variance ratios ρ=3.8 (BBQ), **14.9 [3.2, 249.2]** (ToxiGen), 0.7 (TruthfulQA), 2.4 (XSTest), where the bracketed interval is the within-grid bootstrap range on the ToxiGen headline (App. I). The directional reading — implementation main effects exceed model main effects on three of four benchmarks, with only TruthfulQA model-dominated — is robust; the magnitude on ToxiGen is not. The 14.9 point estimate should be read as evidence of strong amplification, not as a precisely known multiplier: the interval is wide because the model-variance share is small (1.8%, so the ratio is unstable near zero) and interaction shares are large (41–54%). We therefore use ρ as an order-of-magnitude qualitative ordering rather than an identified estimate. Type-III robustness (XSTest drops to ρ=1.3), per-benchmark bootstrap intervals, and a parse-rate ≥ 0.95 validated-subset re-run that preserves the qualitative ordering are reported in Appendix H and Appendix D. See Table 9.

Scoring path as a first-class axis. Switching only the scoring pipeline (free-form regex parse ↔ logprob argmax) moves the score by up to 0.468 on Qwen-2.5-7B / TruthfulQA (per-cell numbers in App. J, Fig. 4); in MTMM terms (Campbell & Fiske, 1959) the two paths score non-equivalent measurands and we report the gap rather than mark either as canonical.

Exploratory note: cross-family scale probe (Qwen-2.5 7B→32B + Yi-1.5 9B→34B). Re-running the conservative-core sub-envelope at the larger sibling in two families (App. Q, Tab. 6) narrows the conservative-core score range to near-zero on ToxiGen, XSTest, and CrowS-Pairs in *both* families, while leaving it large on TruthfulQA and BBQ in both (Qwen-2.5: $0.494\to0.631$ and $0.445\to0.462$; Yi-1.5: $0.217\to0.275$ and $0.411\to0.442$). We report this as a *bounded two-family*

Table 10. TruthfulQA conservative-core ($|C|$=4) item-subset robustness on 3 stratified 80% subsamples (rs43–rs45, n≈236). Per-cell raw-score range across subsamples (pp) plus per-config ordering on (Qwen, Mistral, Yi). All four orderings are invariant across all three subsamples.

config	ordering (rs43–45)	qwen Δ	mistral Δ	yi Δ	cell-flips
T1/free-form	yi > mistral > qwen	1.70	1.28	3.83	0/3
T1/logprob	mistral > qwen > yi	0.43	2.55	0.43	0/3
T3/free-form	qwen ≥ yi > mistral[†]	1.71	1.28	2.98	0/3[†]
T3/logprob	mistral > qwen > yi	0.00	2.13	1.28	0/3
12 cells aggregated		median 1.49 pp, mean 1.63 pp, max 3.83 pp			

[†] T3/free-form, seed 44: Qwen and Yi land at 0.6255 each (exact tie). Listed as qwen ≥ yi > mistral under the deterministic alphabetical tie-break convention used by `rank_flip_stratified.py`; the underlying ordering is a tie, not a strict reversal.

observation on the conservative-core sub-envelope, not a scaling law: it reduces (but does not rule out) the single-family confound on this 4-config envelope, but two families is not evidence of a benchmark taxonomy-by-scale, and we make no claim about Mistral, Llama, or closed-weight models.

W. Disclosure template (GRID card) commentary

Given §5.1 and §U, a benchmark score is reproducible only when accompanied by both its configuration and its harness. The **GRID card** (App. B) is a compact research-level disclosure whose fields cover the implementation axes plus the harness identifier and a minimum statistical summary. It is a starter artefact: no adopter, enforcement, or compliance criteria is specified, and it does not resolve the *standing* problem (a claimant who picks the configuration neighbourhood can game it, as §5.1 shows). Turning configuration disclosure into a deployable verification instrument is institutional work this paper does not do.

X. Item-subset robustness on TruthfulQA conservative core

Robustness to item-subset choice (3 stratified 80% subsamples). The headline is conditioned on a single fixed item subset (n=295 on TruthfulQA, seed 42). To check whether the pairwise-disagreement metric is itself subset-idiosyncratic, we re-ran the 12-config conservative-core slice on three independent stratified 80% subsamples (n≈236 items each, seeds 43–45) on TruthfulQA, yielding 12 (model, configuration) cells × 3 subsamples = 36 raw-score measurements (Tab. 10). The per-cell raw-score range across the three subsamples is small: median 1.49 pp (mean 1.63 pp, max 3.83 pp on Yi/T1/free-form). The within-cell nearest-pair gap is large enough on T1/free-form (every seed ≥ 15 pp) that no plausible ∼2–4 pp resample swing can flip an ordering; for the other three conservative-core configurations the nearest-pair gaps are smaller (mostly 2–8 pp, with one T3/free-form/seed-44 cell where Qwen and Yi land at exactly 0.6255 each — a tie broken by a deterministic alphabetical convention rather than a strict raw-score ordering). The config-level orderings are nevertheless stable: on each of the four conservative-core configurations the induced (Qwen, Mistral, Yi) ordering is the same on every subsample (0/12 cells flip under the deterministic tie convention; the one tied cell is flagged in Tab. 10). ρ_{flip} at $|C|$=4 is discretised in 0.25 increments, so its trivial cross-subsample stability (constant per-pair value: 0.250 / 0.250 / 0.500 for qwen-mistral / qwen-yi / mistral-yi on every subsample) carries less information than the raw-score ranges; the latter is the substantive evidence. We treat this as a conservative-core smoke test on TruthfulQA: no instability was observed across these three 80% subsamples (modulo one Qwen/Yi tied cell that does not contribute a strict ordering on its own), but full-envelope ($|C|$=48) item resampling and resampling on the other four benchmarks remain future work; we therefore do not generalise this stability beyond the 12-cell conservative core on TruthfulQA.

Y. Conservative-core sub-envelope saturation

Robustness to envelope choice. A natural objection to §5.1 is that the envelope itself is hand-curated and could be argued to maximise volatility. We re-compute the same exact rank-flip metric on a *conservative core* sub-envelope that drops every contestable axis value: only templates T1 (minimal) and T3 (canonical lm-eval-harness MCQ); greedy decoding only; 0-shot only; both scoring paths; NF4 (4 configs per free-form benchmark, 2 for CrowS-Pairs). This is the narrowest audited conservative-core sub-envelope inside the core tier — a smoke test, not a representative sample of practice. Table 11 reports the side-by-side metrics.

Table 11. Rank-flip robustness on a conservative-core sub-envelope ($|C|$=4 free-form, $|C|$=2 CrowS). Same metric as Table 2; recall $\rho_{\text{flip}}^{\max}$ is bounded above by 0.5 by combinatorial construction at $|C|$=4 so the "four of five attain 0.5" result is read as "these benchmarks attain the maximum possible value at this envelope size", not as a generic 50% failure rate.

	full envelope		conservative core	
benchmark	$\lvert C\rvert$	$\rho_{\text{flip}}^{\max}$	$\lvert C\rvert$	$\rho_{\text{flip}}^{\max}$
TruthfulQA	48	0.479	4	**0.500**
BBQ	48	0.438	4	**0.500**
ToxiGen	48	0.354	4	**0.500**
XSTest	48	0.292	4	**0.500**
CrowS-Pairs	12	0.250	2	0.000

The conservative-core check therefore evidences *existence* of operator-controllable pairwise reversals even under a strict 4-config sub-envelope, but at $|C|$=4 the metric is upper-bounded by 0.5 by combinatorial construction (any 2–2 split already saturates the bound); the slice cannot estimate the *magnitude* of envelope-driven reversal — only that reversals exist on the narrowest audited sub-envelope. Four of five benchmarks reach this ceiling, meaning that on each of those benchmarks the maximising model pair splits the four conservative-core configurations 2–2. The ceiling is attained for at least one model pair per benchmark, not necessarily for every pair or every cell — even the narrowest audited conservative-core sub-envelope inside the core tier contains an $A{>}B$ and a $B{>}A$ configuration in equal number for the maximising pair. CrowS-Pairs collapses to 0.0 at $|C|$=2 (both configs agree). The full-envelope "5–6 distinct orderings" headline does depend on the broader envelope (T2/T4, few-shot $>$0); the qualitative existence of pairwise reversals on the four free-form benchmarks is preserved on the conservative-core sub-envelope, but the 47.9% TruthfulQA $\rho_{\text{flip}}^{\max}$ magnitude is a full-envelope quantity and should not be inferred from the saturated $|C|$=4 core (where ρ_{flip} takes only the values $0, 0.25, 0.5$).

Z. Claim-scope table

Table 12. Claim scope. Headline numbers in §5.1 sit in the **Supported** column; §U sits in **Suggestive**; everything else is **Out of scope**.

Supported	**Suggestive**	**Out of scope**
Configuration-conditional pairwise-disagreement on the 612-cell NF4 grid (per-benchmark rank-flip envelope: full mixed $\lvert C\rvert$=48, free-form-only $\lvert C\rvert$=36, logprob-only $\lvert C\rvert$=12; CrowS-Pairs $\lvert C\rvert$=12 logprob-only), 3 models $\times$ 5 benchmarks, fixed item subset.	Cross-package score spread at one anchor (illustrative; packages score non-equivalent objects).	Population-level rank stability across model families or model sizes; closed-weight models; jailbreak / multiturn benchmarks.
That an *evaluator inside this envelope* can move pairwise verdicts on up to 47.9% of the $\lvert C\rvert$=48 TruthfulQA harness configurations and produce all 6 orderings on XSTest under the full practice-derived envelope (Tab. 2); on the core-tier envelope (drops T4 and diverse decoding) TruthfulQA $\rho_{\text{flip}}^{\max}$ is 40.7% on $\lvert C\rvert$=27 and XSTest admits 5/6 orderings (Tab. 4).	Cross-family scale-probe collapse on conservative-core (Qwen 7$\rightarrow$32B, Yi 9$\rightarrow$34B; Tab. 6).	A scaling law; a per-axis decomposition of cross-package spread; an enforceable governance instrument.
Variance partition direction (impl. main effects $\geq$ model main effects on three of four free-form benchmarks).	Type-III re-analysis robustness (XSTest is boundary-sensitive).	A precise ρ multiplier (interaction shares 41–54% make ratios qualitative).
GRID card as a research-level disclosure template.	NF4-vs-BF16 score-range comparability on the conservative-core sub-envelope (54 cells; Qwen + Mistral comparable, Yi picks up extra spread on ToxiGen and XSTest).	GRID card as a regulator-ready compliance instrument; a full NF4-vs-BF16 sweep on the full 612-cell grid.