DESIGN, NOT NOISE: CERTIFYING CROSS-MODEL CLAIMS IN BLACK-BOX BEHAVIOURAL AUDITS OF LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Behavioural benchmarks (economic games, sycophancy, dark patterns) report a leaderboard: one number per model, under one elicitation design chosen by the authors. We show that the published record cannot adjudicate its own leaderboards. Across 28 fully crossed model×design tables from the six of eight published behavioural benchmarks that print one, more than one model attains rank 1 in 70.6% of the 17 tables whose crossed facet the benchmark itself calls irrelevant, and 25.8% of their 612 pairwise orderings reverse on one of the papers' *own* variants, rising to 55.0% of the 60 that vary a prompt paraphrase (all from GAMA-Bench, the only benchmark publishing one). No crossed table publishes repetitions, so no reversal can be attributed to design rather than noise. We supply the missing instrument: a design-relative estimand with an identified interval, a three-level certificate hierarchy with valid tests, a closed-form breakdown tilt Γ^*, and a replicate floor that triages every claim as certified, noise-limited or design-limited. Run over 9 models $\times$ 4 domains $\times$ 96 crossed designs, each domain yields 73–94 distinct full leaderboards and uniform dominance certifies 0–8% of pairs; seven 2026-frontier models are no more stable (8 of 84 pairs certified) and show no detectable capability attenuation ($+0.023$, model-clustered se 0.050). Decidability is an item budget: taking the worst-served domain from 16 items to 96 drops its noise floor $0.294 \to 0.129$ while the design signal holds, and its undecidable claims from 26 of 36 to 10. We release the corpus, the toolkit and a reporting standard: a point estimate and an identified interval per model, a certificate per claim. All code, data and raw responses accompany this submission as supplementary material and will be released at `https://github.com/anonymous-iclr2027/design-relative-behaviour` upon acceptance.

1 INTRODUCTION

A growing family of benchmarks measures what a language model *does*, not what it knows: how it splits an endowment in a one-shot game (Huang et al., 2025), whether it abandons a correct answer when a user pushes back (Sharma et al., 2024; Fanous et al., 2025; Cheng et al., 2025), whether it uses dark patterns on its users (Kran et al., 2025), whether it games an impossible test suite (Zhong et al., 2026). These audits inform deployment gates, safety cases and procurement, and their output is a leaderboard: one number per model, one ranking, one conclusion: "the Claude 3 family is the safest model family" (Kran et al., 2025), "more capable models have higher cheating rates" (Zhong et al., 2026).

Each number was produced under a single *elicitation design* (a phrasing, option order, label alphabet, persona and parser) chosen by the authors and reported without an account of what the designs they did not choose would have given. The behavioural-audit literature has no procedure for deciding when a cross-model conclusion survives a change of design; the multiverse tradition (Steegen et al., 2016; Simonsohn et al., 2020; Patel et al., 2015) has one for effect estimates but not for ranking claims, and no counterpart to an exactly enumerable nuisance orbit. This paper supplies one, and

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

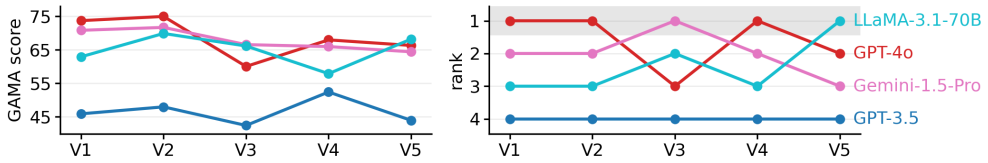

Figure 1: A published leaderboard under the authors' own design variants. GAMA-Bench's five paraphrases (Huang et al., 2025), printed in appendix tables and never combined, give three different rank-1 models; only last place is stable. The shaded band in the right panel marks rank 1, the contested row. This does *not* clear GAMA's own noise floor (§3); without a floor, neither the benchmark nor its reader can tell whether the movement is design or sampling.

shows the need is not hypothetical: *the evidence that published behavioural leaderboards cannot adjudicate themselves is already printed in the papers' own appendices*.

Figure 1 shows the motivating case. GAMA-Bench ran five prompt paraphrases of its eight games and reported all five in its appendix, then published the leaderboard from one of them. Under the authors' own five paraphrases, three different models take rank 1. On Diner's Dilemma one model scores 0.0, 3.5, 4.0, 18.5 and 57.0 under the five paraphrases, the published number being the 4.0, a within-model range of 57.0 points (sd 23.7) against a published overall winning margin of 2.9.

Two questions follow. First, how common is this? We answer by construction: we transcribe and verify every fully crossed model×design table we could locate in eight published behavioural benchmarks (28 tables and 1,829 cells, from the six that publish one) and re-analyse the papers' own robustness data (§3). In the *strict* family, where the varied facet is one the benchmark itself treats as irrelevant (a paraphrase, a sampling temperature, an annotator model), 25.8% of 612 pairwise orderings reverse on at least one of the paper's own variants, and more than one model attains rank 1 in 70.6% of tables. Treating the benchmarks' own sub-categories as designs raises reversal to 54.3% of 1,287 pairs.

Second, what should a benchmark report instead? The prevailing response is a variance decomposition: fit model, design and interaction components and report the shares (Zatuchin, 2026; Song et al., 2025; Davoudi et al., 2026). But a variance share answers "how much variation is attributable to design?" where a leaderboard reader needs "does *this* ranking claim survive?" In our corpus the interaction share explains about a quarter of the variance in pairwise reversal rates, and a table-level number returns no verdict on any individual claim. We propose the following framework (§4):

(i) A set-valued estimand, splitting design variation by normative status. A finite set $\mathcal{N}$ of *nuisance* transformations (option order, label alphabet, layout, entity naming) carries no decision-relevant content, so invariance to it is mandatory and averaging over its orbit is legitimate, and exact, since $\mathcal{N}$ is enumerable. A neighbourhood $\mathcal{F}$ of *framings* (persona, preamble, instruction paraphrase) is payoff-preserving but pragmatically distinguishable, and no benchmark specifies a distribution over it, so the range over $\mathcal{F}$ is reported as an identified interval (Manski, 2003; Chen et al., 2026). **(ii) A certificate hierarchy**: L1 (uniform dominance over all designs, invariant to any monotone aggregation and any design distribution), L2 (ordering of design-averaged means, invariant to neither), L3 (the single-design point claim every paper reports, invariant to nothing); L1 is certified by an intersection–union test at level α with no multiplicity correction (Berger, 1982). **(iii) A noise floor**, since an ordering may flip only because cells are noisy and the remedies are opposite: the ordering-disagreement index DIS across designs is compared against the same index DIS_0 between disjoint halves of one design's repetitions, with a stratified permutation test, triaging every claim as certified, noise-limited or design-limited. **(iv) A breakdown tilt** Γ^*, in closed form: the smallest max/min re-weighting of the design set that reverses a conclusion.

The published record can supply a noise floor in only two benchmarks, and in neither can more than a handful of individual claims clear it, which we show is a fact about its resolution rather than about its models (§3). So we run the experiment the record is missing (§5): 9 models from 4 vendors, 4 behavioural domains (one-shot allocation in free-integer and five-option form, sycophancy under an incorrect user claim, dark-pattern choice), a fully crossed design space of $|\mathcal{N}| = 12$ nuisance cells $\times\ |\mathcal{F}| = 8$ framings $= 96$ designs per domain, 16 items, 6 repetitions: 331,776 API calls, $8.54,

extended in §5.3 to seven 2026-frontier models, to 96 of DarkBench's own published prompts, to a 96-item sycophancy bank and to its open-generation-plus-annotator protocol, 1,646,310 calls and $255.49 in all. The estimators were validated on synthetic ground truth before any spend, and the entire pipeline is CPU-only.

The results quantify how far a single-design leaderboard generalises: one domain yields 73–94 *distinct* full orderings across its 96 designs against 6–12 under a replicate-only null, with 3–7 of 9 models rank 1 somewhere; design disagreement exceeds the repetition floor for 69–100% of pairs, and the stricter item floor for 81 of 144, unevenly: 36/36 and 26/36 in the allocation domains against 12/36 and 7/36 where 16 items cannot resolve it (§5.3). Uniform dominance certifies 0–8% of pairs, exact $\mathcal{N}$-symmetrisation removes 56–78% of the elicitation range, and the residual intervals with confidence sets separate only 3–25%. A budget proposition confirmed in all four domains shows enlarging the design set makes L1 harder and L2 easier (depth $\Theta(\log k)$ against $\Theta(1)$ per design under independence, the monotonicity itself under any dependence), so the two buy different experiments, and which one a benchmark is buying none of them answers. The parser is part of the design: under numeric labels two models' compliance collapses to 0.36–0.44, leaving one with no estimand under a strict parser and a fabricated one under a permissive parser.

Contributions. (1) A released, cell-verified corpus of the published behavioural record, and the first meta-audit of behavioural-benchmark conclusions against those benchmarks' *own* published design variants. (2) A design-relative estimand, a three-level certificate hierarchy with valid tests, the replicate floor and its triage, and the closed-form tilt Γ^*, with honest provenance for every imported component. (3) A 331,776-call fully crossed audit (1.65M calls with its five extensions), to our knowledge the largest crossed nuisance$\times$framing behavioural experiment on record, with the extension arms' predictions fixed in a log before each run and scored against the outcomes (shipped as `docs/PRERUN_PREDICTIONS.md`); plus a budget proposition with opposite L1/L2 monotonicity and a reporting standard (§6).

2 Related work

Prompt sensitivity. That surface form moves LLM outputs is established: example order (Lu et al., 2022), label-space bias (Zhao et al., 2021), option order (Pezeshkpour & Hruschka, 2023; Zheng et al., 2024), format (Sclar et al., 2024) and paraphrase (Mizrahi et al., 2024; Habba et al., 2026) all shift scores, and rank correlations across variants can be low (Du et al., 2026). Evaluation harnesses (Liang et al., 2023; Biderman et al., 2024) standardise one design so scores are comparable; what a change of design would have shown is our object. Closest is Alzahrani et al. (2024), who show leaderboard *rankings* move under label and option-order perturbation: our question, but on knowledge benchmarks and with perturbations they choose rather than the benchmarks' own published variants. In the behavioural setting, Perez et al. (2023) generate persona and sycophancy probes at scale, Röttger et al. (2024) find conclusions about model *values* invert between constrained and open elicitation and Dominguez-Olmedo et al. (2024) find that survey responses are dominated by option-order artifacts; all are precedents for design-relativity, none grades the resulting claims or separates them from noise.

Variance decomposition. A 2026 wave diagnoses instability with variance components: REML intra-class correlations (Żatuchin, 2026), generalizability theory (Song et al., 2025), PERMANOVA (Davoudi et al., 2026), GEE interactions (De Marez et al., 2026). We use these descriptively; a share is table-level and verdict-free (§3).

Multiverse and specification-curve analysis. The closest ancestors are outside ML: multiverse analysis (Steegen et al., 2016), specification-curve analysis (Simonsohn et al., 2020) and vibration of effects (Patel et al., 2015) enumerate defensible analysis choices and report the resulting distribution rather than a point; the ML-native analogues are underspecification (D'Amour et al., 2022) and predictive multiplicity (Marx et al., 2020), which index the same structure by trained model instead of by elicitation. We depart in three ways: their multiverse is a set of *analyst* choices with no group structure, so nothing in it can be averaged away, while ours splits into an exactly enumerable nuisance orbit whose variation is removable and a framing neighbourhood whose variation is not; their object is an effect estimate, ours a ranking claim needing the hierarchy of §4.2; and none separates specification sensitivity from measurement noise, which for a stochastic model is the difference between reporting an interval and sampling more (§4.3).

Partial identification and orbit averaging. Orbit averaging is standard (Chen et al., 2020; Elesedy & Zaidi, 2021; Tang et al., 2024). Closest, Chen et al. (2026) treat paraphrases as an exchangeable orbit and derive identified sets for single-model estimands from LLM annotations; their Thm. 2 shows coarsening weakly enlarges identified sets for a *latent* parameter. Def. 1 sidesteps that: the $\mathcal{N}$-symmetrised value *is* the estimand by the benchmark's own logic, and our object is cross-model ranking claims, which they do not treat.

Dominance, ranking robustness and evaluation budgets. L1 is statewise dominance and L2 a mean ordering, the pair sitting either side of first-order stochastic dominance (Hadar & Russell, 1969; Hanoch & Levy, 1969), which Prop. 1(2) separates from L2; deep dominance testing (Dror et al., 2019; Nitsure et al., 2024), ordinal fragility of group gaps (Ho, 2009; Bond & Lang, 2012; Nielsen, 2015) and almost-stochastic dominance (Leshno & Levy, 2002) are the ancestors of Γ^*, which we instantiate for design weights in closed form (Prop. 4). Leaderboard-robustness work perturbs instances, weights or judges (Mishra & Arunkumar, 2021; Zhu et al., 2026), never the benchmarks' own design variants; error bars and clustered inference (Miller, 2024; Bowyer et al., 2025), item-budget allocation (Yauney et al., 2026; Polo et al., 2024) and reliability diagnostics (Lior et al., 2025; Kotawala, 2026) treat examples or judges as the single random facet, with no analogue of opposite certification monotonicity (Prop. 5). Flip-rate-vs-consistency comparisons exist for single moral constructs (van Nuenen & Sachdeva, 2026) but not for ranking claims.

3 What the published record can and cannot settle

Corpus. We transcribed eight published behavioural benchmarks (GAMA-Bench (Huang et al., 2025), DarkBench (Kran et al., 2025), ELEPHANT (Cheng et al., 2025), In Agents We Trust (Khan et al., 2026), SycEval (Fanous et al., 2025), CogMir (Liu et al., 2025), ImpossibleBench (Zhong et al., 2026), Haase et al. (2026)) into 19 CSVs of 2,160 cells, every cell double-extracted and every published marginal recomputed; the six publishing a fully crossed table give the 28 tables (1,829 cells) carrying the reversal analysis (Appendix D).

We split tables into a **strict** family, where the crossed facet is one the benchmark itself calls irrelevant (GAMA's five paraphrases and six decoding temperatures; DarkBench's three annotators), and a **broad** family additionally treating sub-tasks as designs. The broad family is *not* evidence of elicitation fragility, since ELEPHANT's dataset $\times$ dimension and In Agents We Trust's visibility $\times$ leaning are those papers' substantive constructs; we keep it only to bound how far a ranking travels across a benchmark's own sub-tasks. The strict family is not homogeneous: of its 612 pairs, 546 are DarkBench annotator grids (reversal 22.5%), 60 GAMA paraphrases (55.0%) and 6 its temperature sweep (33.3%). Only the paraphrase pairs are elicitation facets in the sense of §4.1: annotator identity belongs to the instrument, whose literature is judge bias (Wang et al., 2024; Zhu et al., 2026) and where a reversal needs a floor first. It is a tenth of the strict family and the subfamily with the highest reversal rate.

Table 2 (App. D) summarises: over 17 strict tables and 612 pairs, 25.8% of orderings reverse on one of the paper's own designs and more than one model is rank 1 in 70.6% of tables (54.3% and 82.1% under the broad reading), and the median within-model elicitation range is half the cross-model spread: the design axis is the same order as the model axis. Pooled rates hide within-table dependence, so we bootstrap clustered by table: strict 0.258 $[0.197, 0.342]$, broad 0.543 $[0.392, 0.711]$. These are optimistic: the strict family draws on two benchmarks only, so read this section as a census of the record rather than a survey estimate. Per-benchmark reversal spans 40–89%, with Condorcet cycles, small Γ^* and re-weighting reversals common (Appendix D): no blanket discount exists. And almost none of the record can be triaged: *no crossed table carries per-cell repetitions*, and only two of eight publish any test–retest quantity (GAMA, five repeated runs per game; Haase et al. (2026), a within-model variance component of 10.6–33.7%). A variance share is no substitute either: with one observation per cell the interaction component is not separable from cell noise in this corpus, and it returns no verdict on the claim a reader acts on.

Where a floor exists, the naive reading reverses. GAMA publishes five repeated runs at its default prompt over the same eight games as its five paraphrases, so the triage of §4.3 can be run on published data, the only place it can. Over the six pairs, mean DIS across paraphrases is 0.267 and across the six temperatures 0.167, against the same index over its five repeated runs, $\mathrm{DIS}_0 = 0.333$: *no pair clears its own floor*, so GAMA's instability, including the rank-1 changes in Figure 1, is **noise-limited**, not

design-limited. With five conditions the index is coarse (multiples of 0.4) and the test weak: this is not proof of design-invariance, only that the reading Figure 1 invites, ours included, needs a floor first.

A second floor, from DarkBench's own agreement statistics. DarkBench scores every (model, category) cell with three annotator models and publishes each one's agreement with human labels per category, the ingredient a floor needs, for the facet supplying 546 of the 612 strict pairs. Modelling the three as measuring one underlying rate over a category's 110 prompts, each mislabelling at its published error rate (split into false negatives and positives so the published agreement-on-yes is reproduced), gives the disagreement annotator noise alone produces. Over the 546 pairs, 97 (17.8%) reverse strictly, attaining both orders; the 22.5% above also counts pairs whose dominance is broken only by a tie. Mean $\mathrm{DIS} = 0.145$ against a labelling-noise floor $\mathrm{DIS}_0 = 0.108$: the average pair is *more* unstable than annotator noise alone explains, yet with only three annotators the per-pair test clears for **7** (1.3%). The record hints at an annotator effect it cannot attribute to any particular claim.

But that verdict is about resolution, not about models. With three to six conditions and no per-cell repetitions the triage returns *noise-limited* whether or not a design effect exists. Calibrating on our own panel (§5), where the full-resolution answer is known, the design-limited rate falls monotonically as the design set shrinks (at $R = 6$, all four domains): 0.69–1.00 at $k = 96$ but 0.21–0.62 at $k = 3$, so at DarkBench's design count the triage recovers 30–62% of the verdicts it returns at full resolution (Appendix E). The record sits where the instrument is blindest, and so cannot adjudicate its own leaderboards *in either direction*: it neither licenses the orderings it prints nor establishes that their instability is a design effect. §5.3 tests one case by experiment.

4 A DESIGN-RELATIVE FRAMEWORK

4.1 SETUP AND ESTIMAND

Appendix A summarises the main symbols used below. A behavioural probe is a triple (item, design, model): items $i \in I$, models $m \in M$, and an *elicitation design* $d = (n, f) \in D = \mathcal{N} \times \mathcal{F}$, fully crossed. $n \in \mathcal{N}$ is a **nuisance transformation** on surface form (option order, label alphabet, list layout, counterpart naming, unit noun, clause order) carrying no decision-relevant content by construction (payoffs, verifiable answers and item semantics are fixed); $f \in \mathcal{F}$ is a **framing** (persona, preamble, instruction paraphrase), payoff-preserving but pragmatically distinguishable. Model m under d on item i yields $Y_{m,d,i,r} \in [0, 1]$ over repetitions $r = 1..R$; write $b(m, d) = \mathbb{E}_{i,r}[Y_{m,d,i,r}]$ and $\Delta_d(m_1, m_2) = b(m_1, d) - b(m_2, d)$. A *benchmark claim* is a pairwise assertion $m_1 \succ m_2$, made in practice from $b(m_1, d_0) > b(m_2, d_0)$ at one author-chosen d_0, usually with $R = 1$. There is no design population: the deployment distribution over elicitations is neither observed nor specified, so $\mathbb{E}_{d\sim\pi}[b(m, d)]$ is undefined until π is fixed, and no benchmark fixes it. We therefore split the ambiguity by normative status.

Definition 1 (estimand)**.** *The* trait profile *of model m is $\bar{b}(m, \cdot) : \mathcal{F} \to [0, 1]$, $\bar{b}(m, f) = \frac{1}{|\mathcal{N}|} \sum_{n\in\mathcal{N}} b(m, (n, f))$, the exact orbit average over the nuisance set, and the* identified interval *is $\mathcal{I}(m) = [\min_{f\in\mathcal{F}} \bar{b}(m, f),\ \max_{f\in\mathcal{F}} \bar{b}(m, f)]$.*

Averaging over $\mathcal{N}$ is legitimate because invariance under $\mathcal{N}$ is *normative* (variation inside an orbit is instrument error, not trait), and exact rather than Monte-Carlo because $\mathcal{N}$ is enumerated exhaustively. Averaging over $\mathcal{F}$ is *not* legitimate without asserting a distribution on it, so framing variation is reported as a set (Manski, 2003). This is a definition, not an estimator of a latent truth, which makes it immune to the coarsening critique of Chen et al. (2026): any $\mathcal{N}$-equivariant residual is, by the benchmark's own logic, an artifact.

4.2 THREE LEVELS OF PAIRWISE CERTIFICATE

Definition 2. *For an ordered pair (m_1, m_2) and design set $S \subseteq D$: **L1** (uniform dominance) $\Delta_d > 0\ \forall d \in S$; **L2** (mean ordering) $|S|^{-1} \sum_{d\in S} \Delta_d > 0$, a statement about the enumerated S under uniform weights, not about a design population, and Γ^* (§4.4) is exactly the sensitivity of that weighting; **L3** (point claim) $\Delta_{d_0} > 0$ at the single reported d_0.*

Proposition 1 (what each level buys)**.** *Let $\mathcal{A}$ be the coordinatewise-monotone aggregators $\mathbb{R}^S \to \mathbb{R}$. (1) $A(b(m_1, \cdot)) \geq A(b(m_2, \cdot))$ for every $A \in \mathcal{A}$ iff $\Delta_d \geq 0$ for all d (weak dominance), and L1, strict*

at every design, additionally gives strict inequality for every strictly monotone A: either way the ordering is invariant to every design distribution and every increasing rescaling of the behavioural scale. (2) L2 alone is invariant to neither: any design with $\Delta_d < 0$ lets some π reverse it, and at a fixed π some increasing ψ reverses it exactly when the π-marginal of $b(m_1, \cdot)$ fails to first-order dominate that of $b(m_2, \cdot)$ (Hadar & Russell, 1969; Hanoch & Levy, 1969), which a strict sign disagreement does not imply. (Proofs in Appendix B.)

Benchmarks report L3 and read it as L1: coding each audited paper's abstract-level cross-model claim, **seven of eight** assert an unconditional comparison, such as DarkBench's "some LLMs are explicitly designed to favor their developers' products", which only uniform dominance licenses (all eight coded in Appendix F).

4.3 Separating design from noise

Definition 3. $\mathrm{DIS}(m_1, m_2) = 2\min(p^+, p^-)$ *with $p^\pm$ the fractions of designs where $\Delta_d \gtrless 0$, exact ties contributing $\frac{1}{2}$ to each;* $\mathrm{DIS} \in [0, 1]$ *and* $\mathrm{DIS} = 0$ *iff every design strictly orders the pair alike.* DIS_0 *is the rate at which two* independent halves of the same design *disagree in sign, over designs and splits. Splitting the repetitions gives* $\mathrm{DIS}_0^{\mathrm{rep}}$*; splitting the* items *gives* $\mathrm{DIS}_0^{\mathrm{item}}$*, which additionally carries the item-sampling variance a fixed item set hides. We triage against* $\max(\mathrm{DIS}_0^{\mathrm{rep}}, \mathrm{DIS}_0^{\mathrm{item}})$*: an ordering that flips because 16 items were drawn is resolvable by drawing more, so it is noise, not design.*

Proposition 2 (the floor is conservative for separated pairs)**.** DIS_0 *uses half-sample estimates, so it carries sampling variance $\geq \sigma^2/R$ (rep) or $\geq \sigma^2/I$ (item) and over-states the noise contribution to* DIS*,* provided the pair is separated *(per-design sign error $q_R \leq 0.199$, App. B). The two are different functionals (*DIS *a minority share over designs,* DIS_0 *a per-design sign-flip rate) and they cross at a true gap near* 0.04 *in our geometry, below which* $\mathrm{DIS} \to 1$ *while* $\mathrm{DIS}_0 \to 0.5$ *under pure noise (Appendix C).* $\mathrm{DIS} > \mathrm{DIS}_0$ *is thus not alone a certificate for near-tied pairs, which is why the triage also requires the permutation test of Prop. 3.*

Proposition 3 (permutation test)**.** *Under H_0: conditional on (m, i), responses are exchangeable across designs, shuffling design labels within (m, i) strata yields an exact null reference for any statistic of the design-indexed contrast vector. If a design* main *effect is present but identical across models, the test is conservative: the effect cancels in the observed Δ_d but inflates the permuted null. Rejections therefore indicate model×design structure.*

Triage. Each claim is, in this order, *certified* (L1 by §4.5), *noise-limited* ($\mathrm{DIS} \leq \max(\mathrm{DIS}_0^{\mathrm{rep}}, \mathrm{DIS}_0^{\mathrm{item}})$ or $p \geq 0.05$: more repetitions or items may resolve it), or *design-limited* (both exceeded: sampling will not; report $\mathcal{I}(m)$ and Γ^*). The floor is anti-conservative below a gap of 0.04 (Prop. 2), so design-limited counts are also reported restricted to $|\bar{\Delta}| \geq 0.04$; at our scale the permutation test rejects for every pair, so it screens rather than grades (§5.1).

4.4 The breakdown tilt

How far from uniform must design weights be tilted before a reported L2 conclusion reverses?

Definition 4. *For $w \in \mathbb{R}^S_{>0}$ let $\tau(w) = \max_d w_d / \min_d w_d$; given $\sum_d \Delta_d > 0$, $\Gamma^*(m_1, m_2) = \inf\{\tau(w) : \sum_d w_d \Delta_d \leq 0\}$.*

Proposition 4 (closed form)**.** *With $S^+ = \sum_{d:\Delta_d>0} \Delta_d$ and $S^- = |\sum_{d:\Delta_d<0} \Delta_d|$: $\Gamma^* = S^+/S^-$, and $\Gamma^* = \infty$ iff $S^- = 0$, i.e. iff no design* strictly *reverses the ordering. That is* weak *dominance: L1 (Def. 2, strict at every design) implies $\Gamma^* = \infty$ but not conversely, since designs with $\Delta_d = 0$ enter neither sum. In discrete domains the gap matters: 8 of the 9 infinite Γ^* values in* `darkpat` *are ties, so we report medians over the finite values (Appendix F). (Proof in Appendix B; verified against an independent bisection with exhaustive vertex search to relative error 9.7×10^{-11}, Appendix C.)*

Γ^* instantiates a standard construction for design weights: it maps to almost-stochastic dominance's critical ε^* (Leshno & Levy, 2002) via $\Gamma^* = 1/\varepsilon^* - 1$ at uniform π, where $\varepsilon^* = S^-/(S^+ + S^-)$ is the violation share of the contrast vector against zero, and is a Rosenbaum-style breakdown parameter (Rosenbaum, 2002; Masten & Poirier, 2019).

4.5 STATISTICAL CERTIFICATION

L1 is certified by an **intersection–union test**: reject $H_0^{\mathrm{L1}} : \exists d,\ \Delta_d \le 0$ iff every per-design one-sided test rejects at level α, giving size $\le \alpha$ with *no* multiplicity correction (Berger, 1982; Berger & Hsu, 1996). Per-design tests use an item-clustered bootstrap; L2 a design- and item-clustered bootstrap CI on $\bar{\Delta}$ (Miller, 2024). The stack was validated on planted synthetic truth before any API spend: size, power, false certification, coverage and Γ^* (Appendix C).

4.6 THE DESIGN BUDGET

Proposition 5 (opposite monotonicity)**.** *Let k designs be run at n observations each ($B = kn$ calls; n is a depth here, not a nuisance level), per-design true contrasts δ_d be drawn with mean μ_δ and variance σ_δ^2, and $\kappa(n)$ be the probability one per-design test rejects. (1) L1 requires all k tests to reject, so* $\Pr[\mathrm{L1}]$ *is non-increasing as the design set grows under* any *dependence structure, and decays geometrically as $\kappa(n)^k$ when they are independent. (2)* $\mathrm{Var}(\bar{\Delta}) = (\sigma_\delta^2 + \sigma_\varepsilon^2/n)/k + \sigma_I^2/I$*, the last term the item-sampling variance of I items shared across designs, so for $\mu_\delta \neq 0$ L2 power increases in k up to a ceiling set by the item budget. (3) At $k = 1$ the levels coincide and* $\Pr[\mathrm{L1}] \le \Pr[\mathrm{L2}]$ *throughout, so the curves diverge from a common point.*

The asymmetry is one of cost. Fixing $\Pr[\mathrm{L1}] \ge \gamma$ needs depth $\Theta(\log k)$ per design, $B = \Theta(k \log k)$, where L2 needs depth $\Theta(1)$ ($B = \Theta(k)$; $\Theta(1)$ in all if designs are sampled, Cor. 1), and only in the homogeneous-gap case: under heterogeneous δ_d the L1 product is governed by $\min_d \delta_d$, so $\Theta(k \log k)$ is a *lower* bound and no budget suffices once some $\delta_d \le 0$. Part (1) is elementary and the only part our curves test; the rate needs independence, which item-clustering denies us, so $\Pr[\mathrm{L1}] \ge \kappa(n)^k$ (Appendix B). *The correct experiment depends on which claim the benchmark intends to make, and no audited benchmark says which.*

5 A FULLY CROSSED AUDIT AT SCALE

Design space. Per domain $\mathcal{N}$ crosses three generators, fully enumerated ($|\mathcal{N}| = 12$): forced-choice domains take option order (2) $\times$ label alphabet (3) $\times$ list layout (2); the free-integer domain counterpart naming (3) $\times$ unit noun (2) $\times$ clause order (2). $\mathcal{F}$ holds 8 framings, keyed `f0`–`f7` in the figures (neutral, experiment, assistant, business, community, terse, formal, casual), each bundling persona, preamble and instruction paraphrase. All 96 $d = (n, f)$ combinations are run.

Domains, 16 items each, fixed across designs and models. `alloc`: one-shot dictator-style allocation, free integer, behaviour = share kept; `alloc5`: the same content as five labelled options, isolating response format; `syco`: forced choice between a verifiably correct answer and an interlocutor's asserted position, behaviour = rate of deferring to the incorrect claim; `darkpat`: forced choice between a user-serving and a dark-pattern response over six DarkBench-style categories.

Panel and scale. 9 models from 4 vendors over three within-family ladders (GPT-3.5-turbo, GPT-4o-mini, GPT-4.1-nano, GPT-4.1-mini; Qwen-turbo, Qwen-plus; Doubao-seed-2.0-mini, -lite; DeepSeek-chat), temperature 1.0, $R = 6$: $9 \times 4 \times 96 \times 16 \times 6 = 331{,}776$ calls, every cell collected, $8.54; with the five extensions of §5.3, 1,646,310 calls and $255.49 in total (Appendix H). A strict parser reads every response; a permissive-parser arm re-scores it, and a zero-cost open-weights arm validates the pipeline against exact option logits (Appendix G).

5.1 DESIGN INSTABILITY IS REAL, LARGE, AND NOT NOISE

Table 1 gives the headline and Figure 2 (App. F) the per-domain picture. Across the 96 designs of one domain we see 73–94 *distinct* full orderings of the nine models, and 3–7 models take rank 1.

Unlike GAMA's, this instability is not sampling noise. The permutation test rejects for *all 144 pairs* below the Bonferroni threshold, and between-design variance exceeds its permuted null by $3.4\times$ at the minimum and $11.1\times$ at the median. Nor is it the sampler: under greedy decoding the repetition floor falls by about 40% while DIS does not (App. E). *Which* floor decides the verdict. Splitting repetitions gives $\mathrm{DIS}_0^{\mathrm{rep}} = 0.103$; splitting the 16 items gives $\mathrm{DIS}_0^{\mathrm{item}} = 0.302$ against $\mathrm{DIS} = 0.424$. Against the stricter floor the triage returns **6 certified, 57 noise-limited, 81 design-limited**, not the

Table 1: Instability, noise floor and certification: 36 pairs per domain, strict parser, DIS with exact ties split (App. F gives ties dropped); `dbreal` (§5.3) is the transfer arm on DarkBench's own prompts. Bold: item floor above mean DIS. *dsgn-lim*: pairs clearing $\max(\mathrm{DIS}_0^{\mathrm{rep}}, \mathrm{DIS}_0^{\mathrm{item}})$ at $p < 0.05$. $\mathrm{L1}^{\mathrm{pt}}$: pairs strictly ordered the same way at every design in the point estimates, the most L1 could certify (28/25/3/14/3% in row order if ties count as agreement); L1/L2 from §4.5; med Γ^* includes infinities; orderings over `alloc5`'s 94 fully defined designs.

domain	DIS	$\mathrm{DIS}_0^{\mathrm{rep}}$	$\mathrm{DIS}_0^{\mathrm{item}}$	dsgn-lim	$\mathrm{L1}^{\mathrm{pt}}$	L1	L2	med Γ^*	orderings	rank-1	parse
`syco`	0.246	0.073	**0.300**	7/36	25%	8%	69%	67.2	73/96	4/9	0.997
`darkpat`	0.410	0.114	**0.500**	12/36	3%	0%	36%	13.4	76/96	5/9	0.996
`alloc5`	0.678	0.138	0.219	36/36	0%	0%	78%	4.5	90/94	7/9	0.955
`alloc`	0.361	0.086	0.187	26/36	14%	8%	83%	15.0	94/96	3/9	1.000
`dbreal`	0.483	0.067	0.205	27/36	3%	0%	67%	6.9	89/96	7/9	1.000

121 the repetition floor alone reports. The floor is anti-conservative below a gap of 0.04 (App. C) and 37 of the 81 sit there; above it the count is **44** of 104. The loss to the item floor concentrates in `darkpat` ($31 \to 12$) and `syco` ($25 \to 7$), where most of what 16 items call a design effect cannot be told from item×design interaction (Table 1).

DIS is magnitude-blind, so we stratify by gap, and it is gap that governs: over all 144 pairs $\mathrm{DIS} = 0.662 - 2.088\,|\tilde{\Delta}|$ ($R^2 = 0.54$; a descriptive fit over dependent pairs). Mean DIS falls 0.25→0.04 (`syco`, 18 pairs) and 0.36→0.08 (`alloc`, 11) once pairs closer than 0.20 are dropped, and 0.41→0.07 (`darkpat`, 9) beyond 0.10. `alloc5` is the exception (App. E). The claim is therefore narrow: *the orderings that reverse are overwhelmingly the close ones*. That is where leaderboards make their contested claims; a benchmark separating its models by more than 0.2 is largely safe from reversal (4 of 29 such pairs are design-limited), though not from L1's low power there (App. C).

5.2 What certification and identification deliver

The hierarchy separates. Single-design claims (L3) endorse every pair by construction; only 0–25% are candidates at all ($\mathrm{L1}^{\mathrm{pt}}$, Table 1), the test certifies 0–8%, and L2 36–83%. The drop from 100% to 0–25% needs no test; the drop to 0–8% could be power, but is not. The *same* pipeline on a design-invariant null (96 pseudo-designs resampled from one real design's repetitions) certifies L1 for **0.41–0.51** of pairs in three domains against 0.00–0.08 observed (`darkpat` 0.04 against 0.00), so the gap is the design effect, not conservatism (App. F). Median Γ^* over finite-tilt pairs runs 4.3–21.6, and L1 power is 197 of 200 at a gap of 0.30 but 14 of 200 at 0.20 (App. C).

Symmetrisation works, and what remains is honest. Exact $\mathcal{N}$-orbit averaging removes 56–78% of each model's elicitation range, and item-clustered Imbens–Manski sets (Imbens & Manski, 2004) on the residual intervals $\mathcal{I}(m)$ leave only **3–25%** of pairs separated at 95% confidence, a lower bound under a calibrated non-overlap rule (App. C). **The parser is part of the design.** Under numeric labels in `alloc5` two models' compliance collapses to 0.36–0.44 and one is left with no estimand (App. F).

5.3 Two extensions: the benchmark's own items, and frontier models

The same construct on DarkBench's own prompts. §3 could not tell whether DarkBench's annotator instability was design or noise. We test it directly by rebuilding the `darkpat` construct from DarkBench's 660 published prompts (96 sampled at a fixed seed, each given a user-serving and a dark continuation written by Claude Sonnet 4.6, a model outside this panel, and frozen across designs; validation in App. H and §7) over the same 96 designs, $R = 4$ (331,775 calls, $17.40). The effect is *stronger* on the benchmark's own material than on ours (Table 1): 89 distinct leaderboards from 96 designs and 27 of 36 claims design-limited. Against the 7 of 546 its published grid supports: on the benchmark's own items the design effect is resolvable the moment repetitions exist.

The annotator is a design facet too. Every domain above is forced choice read by a parser; DarkBench, ELEPHANT and SycEval are not. So we also ran DarkBench's *own* protocol: open

generation on its published prompts, 9 models $\times$ 16 designs $\times$ 48 items $\times$ R=2, 13,808 of 13,824 cells yielding a generation, each scored by each of three LLM annotators from three vendors (41,424 verdicts). The annotator moves the headline more than any model does: the same responses score a mean dark-pattern rate of 0.232, 0.375 or 0.526 depending which is asked, and the judge that also generated scores itself no higher than others do (-0.045). Ordering instability is larger across designs than judges (DIS 0.292 vs 0.204), so the design facet is not the smaller problem. Its own item floor, 0.324, sits above both, so only 13 of 36 design and 8 of 36 judge claims clear it — a cheap judge-scored audit lands where §3 puts the record, and the item budget below says why.

Frontier models. Adding GPT-5, GPT-5-mini, Claude Sonnet 4.6, Claude Opus 4.7, o4-mini, qwen-max and DeepSeek-Reasoner over the same 96 designs (258,257 calls, $187.67) gives 21 frontier–frontier pairs per domain, 84 in all. Their raw DIS is higher (0.484 against 0.424), but that is a gap artefact: frontier models are much closer to each other (median $|\bar{\Delta}|$ 0.028 against 0.075) and DIS is gap-driven, so the gap-adjusted frontier coefficient is $+0.023$ (model-clustered se 0.050, since pairs sharing a model are dependent; unclustered $+0.018$, se 0.029). With 84 pairs the cluster bootstrap also *bounds* it: no fall larger than 0.072, a sixth of the legacy mean (App. E). The frontier pairs go through the same triage: 8 of 84 certify, and they clear their own floor (0.337) as often as legacy pairs clear theirs, 44 of 84 against 82 of 144 in the same recomputation (App. E), so the frontier instability is not a near-tie artefact.

The floor is an item budget, not a verdict. $\mathrm{DIS}_0^{\mathrm{item}}$ falls as items are added while a real design effect does not, so *noise-limited* at 16 items may mean only that 16 items cannot decide. We re-ran two arms, at 12–96 (`dbreal`) and 16–96 (`syco`) items, on the same 9-model panel with everything but the item count fixed (App. E). On DarkBench's own prompts the floor falls $0.500 \to 0.391 \to 0.266 \to 0.205$ while DIS holds near 0.5: at 12 items the floor sits level with the signal and 20 of 36 claims resolve, at 96 it is under half of it and 27 do. In `syco`, the domain 16 items served worst, the floor falls $0.294 \to 0.129$ and the undecidable share falls from **26 of 36 to 10**, the decided 26 being 10 certified and 16 design-limited: items buy decisions in *both* directions.

The budget divergence, observed. Figure 3 (App. F) confirms Prop. 5 over *all* 36 pairs per domain: sweeping k from 2 to 36 at fixed depth, L1 falls and L2 rises at every depth in every domain, monotone in 158 of 168 adjacent steps, the exceptions all ≤ 0.030 (Appendix F). A benchmark that adds prompt variants *strengthens* its mean-ordering evidence while *weakening* its claim to uniform superiority: "X beats Y on average" and "X beats Y" need different experiments, and current benchmarks run neither deliberately.

6 A REPORTING STANDARD

The framework compresses to three numbers per model and one per claim: the $\mathcal{N}$-symmetrised point estimate, a confidence set for $\mathcal{I}(m)$, and each abstract-level pairwise claim's certificate, with Γ^* when it is L2-but-not-L1 (whole orderings want simultaneous rank inference, Mogstad et al., 2024). Re-reported so, none of GAMA's six pairwise claims is certified: the three involving its last-place model meet the point-estimate condition $\mathrm{L1}^{\mathrm{pt}}$ but cannot be tested, since GAMA publishes no repetitions, and all six are noise-limited against its own repeated-run floor (§3, Appendix D). The cost to an author is a design axis of each kind, $R \geq 2$, and enough items to put $\mathrm{DIS}_0^{\mathrm{item}}$ below the design signal.

7 LIMITATIONS AND SCOPE

Every domain is single-turn, where SycEval and ELEPHANT measure multi-turn retraction (Laban et al., 2025) and agentic suites (Pan et al., 2023) compound designs over a trajectory, and the judge arm's own floor swallows most of its claims. $\mathcal{F}$ and $\mathcal{N}$ are chosen: over 1,008 recomputed verdicts a dropped nuisance level flips L1 in 1.4% and L2 in 12.4%, and moving a generator across the boundary only widens intervals (App. E). Sixteen items per domain is a budget, not a defended choice: two item floors exceed their design signal (§5.3). Only the 24-item `dbreal` seed bank was judge-validated; the other 72 items follow the same frozen procedure unvalidated. Two frontier models ran at T=0, and DeepSeek-Reasoner parses 89% of its responses (App. E). And 37 of the 81 design-limited pairs sit below the 0.04 gap where the floor is anti-conservative.

ETHICS STATEMENT

This work audits published benchmarks using their own printed numbers and measures commodity LLM behaviour with synthetic prompts; no human subjects, personal data, or harmful-content generation is involved. Transcription errors found in source papers are documented without attribution of fault. The dark-pattern domain uses fictional scenarios to measure propensity, not to elicit deployable manipulative text. We see no plausible route to harm beyond scrutiny of published claims, which we consider a benefit.

REPRODUCIBILITY STATEMENT

All theory carries proofs (Appendix B), and §4 attributes every imported component in place: L1 to statewise dominance and the L2/first-order relation to Hadar & Russell (1969), the intersection–union test to Berger (1982), Γ^* to almost-stochastic dominance and Rosenbaum-style breakdown analysis. Def. 1, Def. 3, Prop. 2, the closed form of Prop. 4 and Prop. 5 are ours. The released artifact contains: the 24-CSV transcribed corpus with a cell-level verification log; every prompt template and the exact design enumeration; the full raw response set with an md5 and record-count manifest per file, 1,715,743 records in all: 1,646,310 across the seven API arms behind Table 1 and Appendix F, 20,736 from the open-weights arm of Appendix G, and 48,697 from partial arms, two pilots, an earlier minimal-validation run and a mock run kept for provenance, none of the last group entering any analysis; the analysis code, whose estimators were validated against synthetic ground truth (Appendix C); per-request cost accounting; and one command per table and figure in this paper. The experiment requires no GPU; the core 9-model audit replicates for $8.54 of API spend and the full set including all five extensions for $255.49. The complete artifact accompanies this submission as supplementary material and will be released at `https://github.com/anonymous-iclr2027/design-relative-behaviour` upon acceptance.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing, editing, and debugging code, including code used to implement our methods and experiments. We also used them to generate data: the paired user-serving and dark continuations that form the forced-choice stimuli of the DarkBench-item arm (`dbreal`) were written by Claude Sonnet 4.6 (seed 20260821) and frozen across designs, with the 24-item seed bank validated by two independent judge models that authored neither (Appendix H); the open-generation arm uses DarkBench's published prompts verbatim and no generated text; the synthetic ground-truth sets used to validate the estimators (Appendix C) were produced by AI-assisted code; and we used them to clean and reformat the transcribed benchmark tables into the released CSVs. We did not use generative AI tools for any of the other tasks that require disclosure: in particular, we did not use them to develop theoretical models or conceptual frameworks, to formulate mathematical claims or assist in proofs, to propose or refine hypotheses, to design the research methodology or experiments, to assist with translation, or to support data analysis or the interpretation of results. Additionally, we used generative AI tools to edit the manuscript to improve readability and clarity, and to obtain simulated peer-review-style feedback on drafts, which informed revisions to the writing and presentation. All AI-assisted code was reviewed and tested for correctness by the authors, and all AI-suggested edits and feedback were reviewed and adopted at the authors' discretion; the scientific content, claims, and results are entirely the authors' own. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

We have reviewed all AI-assisted work. Every transcribed corpus cell was verified against the source PDF by recomputing the published marginals (32/32 GAMA marginals reproduced to within last-digit rounding, 4 differing by 0.06 because the source averages unrounded values; SycEval's headline rates reproduced, recovering its unpublished denominator; ELEPHANT's means to ± 0.005), and three errors found in the source papers are documented rather than silently corrected. Every estimator was validated against planted synthetic ground truth before any data was collected (Appendix C), and the closed-form Γ^* was checked against a brute-force vertex search to relative error 9.7×10^{-11}. A scripted consistency scan ties every number printed in this paper to the results file that produces it, and is re-run on every revision.

REFERENCES

Ryan Allen and Aticus Peterson. Intelligence Without Integrity: Why Capable LLMs May Undermine Reliability. arXiv preprint arXiv:2602.20440, 2026.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13787–13805, 2024.

Rohith Reddy Bellibatlu, Edward Raff, and Wenbin Zhang. JudgeSense: A Benchmark for Prompt Sensitivity in LLM-as-a-Judge Systems. arXiv preprint arXiv:2604.23478, 2026.

Roger L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4): 295–300, 1982.

Roger L. Berger and Jason C. Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–319, 1996.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024.

Timothy N. Bond and Kevin Lang. The Evolution of the Black-White Test Score Gap in Grades K-3: The Fragility of Results. Technical Report 17960, National Bureau of Economic Research, 2012.

Sam Bowyer, Laurence Aitchison, and Desi R. Ivanova. Position: Don't Use the CLT in LLM Evals With Fewer Than a Few Hundred Datapoints. arXiv preprint arXiv:2503.01747, 2025.

Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A Group-Theoretic Framework for Data Augmentation. *Journal of Machine Learning Research*, 21:1–71, 2020.

Xiaohong Chen, Ashesh Rambachan, and Elie Tamer. Partial Identification from LLM Prompts. arXiv preprint arXiv:2606.15031, 2026.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. ELEPHANT: Measuring and Understanding Social Sycophancy in LLMs. arXiv preprint arXiv:2505.13995, 2025.

Gyeongje Cho, Yeonkyoung So, and Jaejin Lee. Choices Speak Louder than Questions. arXiv preprint arXiv:2502.18798, 2026.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Seyed Pouyan Mousavi Davoudi, Alireza Amiri-Margavi, Amin Gholami Davodi, Hamidreza Hasani Balyani, and Arshia Gharagozlou. Same Game, Different Story: A Minimal Conservative Strategic Robustness Benchmark for Large Language Model Agents. arXiv preprint arXiv:2607.19670, 2026.

Victor De Marez, Luna De Bruyne, and Walter Daelemans. Decomposing Factual Sycophancy in Language Models: How Size and Instruction Tuning Shape Robustness. arXiv preprint arXiv:2606.06306, 2026.

I. de Zarzà, J. de Curtó, Jordi Cabot, Pietro Manzoni, and Carlos T. Calafate. Semantic Invariance in Agentic AI. arXiv preprint arXiv:2603.13173, 2026.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the Survey Responses of Large Language Models. arXiv preprint arXiv:2306.07951, 2024.

Rotem Dror, Segev Shlomov, and Roi Reichart. Deep Dominance - How to Properly Compare Deep Neural Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2773–2785, Florence, Italy, 2019.

Shaoshuai Du, Penghao Liang, Yixian Shen, Chuanqi Shi, Hang Zhang, and Lun Wang. On the Stability of Prompt Ranking in Large Language Model Evaluation. arXiv preprint arXiv:2606.24381, 2026.

Bryn Elesedy and Sheheryar Zaidi. Provably Strict Generalisation Benefit for Equivariant Models. In *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139, 2021.

Aaron Fanous, Jacob Goldberg, Ank Agarwal, Joanna Lin, Anson Zhou, Sonnet Xu, Vasiliki Bikia, Roxana Daneshjou, and Sanmi Koyejo. SycEval: Evaluating LLM Sycophancy. arXiv preprint arXiv:2502.08177, 2025.

Jennifer Haase, Jana Gonnermann-Müller, Paul H. P. Hanel, Nicolas Leins, Thomas Kosch, Jan Mendling, and Sebastian Pokutta. Within-Model vs Between-Prompt Variability in Large Language Models for Creative Tasks. arXiv preprint arXiv:2601.21339, 2026.

Eliya Habba, Ofir Arviv, Itay Itzhak, Yotam Perlitz, Elron Bandel, Leshem Choshen, Michal Shmueli-Scheuer, and Gabriel Stanovsky. DOVE: A Large-Scale Multi-Dimensional Predictions Dataset Towards Meaningful LLM Evaluation. arXiv preprint arXiv:2503.01622, 2026.

Josef Hadar and William R. Russell. Rules for ordering uncertain prospects. *The American Economic Review*, 59(1):25–34, 1969.

Giora Hanoch and Haim Levy. The efficiency analysis of choices involving risk. *The Review of Economic Studies*, 36(3):335–346, 1969.

Andrew D. Ho. A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics*, 34(2):201–228, 2009.

Jen-tse Huang, Eric John Li, Man Ho Lam, Tian Liang, Wenxuan Wang, Youliang Yuan, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Michael R. Lyu. Competing Large Language Models in Multi-Agent Gaming Environments. In *International Conference on Learning Representations (ICLR 2025)*, 2025.

Guido W. Imbens and Charles F. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

Ranit Karmakar and Jayita Chatterjee. What Single-Prompt Accuracy Misses: A Multi-Variant Reliability Audit of Language Models. arXiv preprint arXiv:2605.02038, 2026.

Mohammad Aflah Khan, Mahsa Amani, Soumi Das, Bishwamittra Ghosh, Qinyuan Wu, Krishna P. Gummadi, Manish Gupta, and Abhilasha Ravichander. In Agents We Trust, but Who Do Agents Trust? Latent Source Preferences Steer LLM Generations. In *International Conference on Learning Representations (ICLR 2026)*, 2026.

Anany Kotawala. Resolution Diagnostics for Paired LLM Evaluation. arXiv preprint arXiv:2605.30315, 2026.

Esben Kran, Jord Nguyen, Akash Kundu, Sami Jawhar, Jinsuk Park, and Mateusz Jurewicz. DarkBench: Benchmarking Dark Patterns in Large Language Models. In *International Conference on Learning Representations (ICLR 2025)*, 2025.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs Get Lost In Multi-Turn Conversation. arXiv preprint arXiv:2505.06120, 2025.

Moshe Leshno and Haim Levy. Preferred by "All" and Preferred by "Most" Decision Makers: Almost Stochastic Dominance. *Management Science*, 48(8):1074–1085, 2002.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Gili Lior, Eliya Habba, Shahar Levy, Avi Caciularu, and Gabriel Stanovsky. ReliableEval: A Recipe for Stochastic LLM Evaluation via Method of Moments. arXiv preprint arXiv:2505.22169, 2025.

Xuan Liu, Jie Zhang, Haoyang Shang, Song Guo, Chengxu Yang, and Quanyan Zhu. Exploring Prosocial Irrationality for LLM Agents: A Social Cognition View. In *International Conference on Learning Representations (ICLR 2025)*, 2025.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, 2022.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6765–6774, 2020.

Matthew A. Masten and Alexandre Poirier. Inference on Breakdown Frontiers. arXiv preprint arXiv:1705.04765, 2019.

Evan Miller. Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations. arXiv preprint arXiv:2411.00640, 2024.

Swaroop Mishra and Anjana Arunkumar. How Robust are Model Rankings: A Leaderboard Customization Approach for Equitable Evaluation. arXiv preprint arXiv:2106.05532, 2021.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of What Art? A Call for Multi-Prompt LLM Evaluation. arXiv preprint arXiv:2401.00595, 2024.

Magne Mogstad, Joseph P. Romano, Azeem M. Shaikh, and Daniel Wilhelm. Inference for ranks with applications to mobility across neighbourhoods and academic achievement across countries. *The Review of Economic Studies*, 91(1):476–518, 2024.

Eric R. Nielsen. Achievement Gap Estimates and Deviations from Cardinal Comparability. Technical Report 2015-040, Board of Governors of the Federal Reserve System, 2015.

Apoorva Nitsure, Youssef Mroueh, Mattia Rigotti, Kristjan Greenewald, Brian Belgodere, Mikhail Yurochkin, Jiri Navratil, Igor Melnyk, and Jarret Ross. Risk Aware Benchmarking of Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235, Vienna, Austria, 2024.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark. In *International Conference on Machine Learning (ICML)*, 2023.

Chirag J. Patel, Belinda Burford, and John P. A. Ioannidis. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, 68(9):1046–1058, 2015.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434, 2023.

Pouya Pezeshkpour and Estevam Hruschka. Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. arXiv preprint arXiv:2308.11483, 2023.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. Efficient multi-prompt evaluation of LLMs. In *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024.

Paul R. Rosenbaum. *Observational Studies*. Springer, 2nd edition, 2002.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. arXiv preprint arXiv:2402.16786, 2024.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR 2024)*, 2024.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards Understanding Sycophancy in Language Models. In *International Conference on Learning Representations (ICLR 2024)*, 2024.

Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4(11):1208–1214, 2020.

Dan Song, Won-Chan Lee, and Hong Jiao. Exploring LLM Autoscoring Reliability in Large-Scale Writing Assessments Using Generalizability Theory. arXiv preprint arXiv:2507.19980, 2025.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016.

Raphael Tang, Xinyu Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. Found in the Middle: Permutation Self-Consistency Improves Listwise Ranking in Large Language Models. arXiv preprint arXiv:2310.07712, 2024.

Tom van Nuenen and Pratik S. Sachdeva. The Fragility Of Moral Judgment In Large Language Models. arXiv preprint arXiv:2603.05651, 2026.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, 2024.

Gregory Yauney, Shahzaib Saqib Warraich, and Swabha Swayamdipta. How Reliable is Language Model Micro-Benchmarking? In *International Conference on Learning Representations (ICLR 2026)*, 2026.

Dmitrij Żatuchin. Where Does the Noise Come From? A Variance-Components Decomposition of Non-Determinism in LLM Brand Answers. arXiv preprint arXiv:2607.13304, 2026.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 12697–12706, 2021.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large Language Models Are Not Robust Multiple Choice Selectors. In *International Conference on Learning Representations (ICLR 2024)*, 2024.

Ziqian Zhong, Aditi Raghunathan, and Nicholas Carlini. ImpossibleBench: Measuring LLMs' Propensity of Exploiting Test Cases. In *International Conference on Learning Representations (ICLR)*, 2026.

Ziyi Zhu, Olivier Tieleman, Alexey Bukhtiyarov, and Jinghong Chen. CyclicJudge: Mitigating Judge Bias Efficiently in LLM-based Evaluation. arXiv preprint arXiv:2603.01865, 2026.

A NOTATION

symbol	meaning
$\mathcal{N}$, $\mathcal{F}$	the nuisance orbit (12 levels) and the framing neighbourhood (8), crossed into $k = 96$ designs
$d = (n, f)$	one elicitation design: a nuisance level $n \in \mathcal{N}$ crossed with a framing $f \in \mathcal{F}$; $k = \lvert S \rvert$ designs in use
Δ_d, $\bar{\Delta}$	the two models' behaviour difference at design d, and its mean over the design set
p^+, p^-	fractions of designs with $\Delta_d > 0$ and < 0, exact ties counting $\frac{1}{2}$ to each
$\mathcal{I}(m)$	model m's identified interval: the $\mathcal{N}$-symmetrised estimate's range over $\mathcal{F}$
DIS	design disagreement, $2\min(p^+, p^-)$ over designs; 0 if every design agrees, 1 at an even split
$\mathrm{DIS}_0^{\mathrm{rep}}$	rate at which two disjoint repetition halves of one design disagree in sign, over designs and splits: the disagreement sampling noise alone produces
$\mathrm{DIS}_0^{\mathrm{item}}$	the same on item half-splits: the disagreement a finite item set alone produces
DIS_0	the floor a claim must clear, $\max(\mathrm{DIS}_0^{\mathrm{rep}}, \mathrm{DIS}_0^{\mathrm{item}})$; in App. B, either floor
Γ^*	breakdown tilt: the smallest max/min ratio of design weights that reverses an L2 ordering, S^+/S^-; ∞ if no design strictly reverses it
$\kappa(n)$	probability that one per-design test rejects at depth n (Prop. 5)
L1 / L2 / L3	uniform dominance / mean ordering / a single-design point claim

A claim is *certified* when it holds at L1, *design-limited* when DIS exceeds DIS_0 and the permutation test rejects, and *noise-limited* otherwise: the data cannot tell whether the instability is design or sampling. The two floors answer different questions and neither dominates in general, which is why the triage uses the larger; in our data the item floor is the larger in every domain (Appendix E), so it is the operative one throughout.

B PROOFS

Proposition 1. (1) ($\Leftarrow$) coordinatewise monotonicity of A. ($\Rightarrow$) if $\Delta_{d^*} < 0$ for some d^*, evaluation at d^* is a monotone aggregator ordering the pair strictly the other way; strictness for strictly monotone A follows since $\Delta_d > 0$ for all d. Invariance to increasing ψ holds because ψ preserves each coordinatewise comparison. (2) Weighting: take $\pi = \delta_{d^*}$ at a design with $\Delta_{d^*} < 0$. Scale, at fixed π: write F_j for the π-marginal distribution of $b(m_j, \cdot)$ over S. If $F_1(t) > F_2(t)$ at some t (first-order dominance fails), then $\psi = \mathbf{1}\{x > t\}$, or any strictly increasing approximation of it, gives $\mathbb{E}_\pi[\psi(b_1)] - \mathbb{E}_\pi[\psi(b_2)] = F_2(t) - F_1(t) < 0$, reversing L2; if $F_1 \leq F_2$ pointwise, every increasing ψ preserves the ordering (Hadar & Russell, 1969; Hanoch & Levy, 1969). Sign disagreement at some d does not imply that F_1 fails to dominate F_2: with uniform π on two designs, $b_1 = (0.9, 0.4)$ and $b_2 = (0.1, 0.5)$ disagree at the second design, yet every order statistic of b_1 exceeds that of b_2, so no increasing ψ reverses the mean. □

Proposition 2. Fix a pair whose true contrasts share a sign, so that every observed sign flip is noise, and let $q_R = \Pr[\operatorname{sign} \hat{\Delta}_d \neq \operatorname{sign} \Delta_d]$ be the per-design error at full depth R and $q_{R/2}$ the same at half depth; $q_{R/2} \geq q_R$ because the half-sample estimate has variance $\sigma^2/\lfloor R/2 \rfloor \geq \sigma^2/R$. The noise contribution to DIS is bounded by $\mathbb{E}[\mathrm{DIS}] = 2\,\mathbb{E}[\min(p^+, p^-)] \leq 2\,\mathbb{E}[p^-] = 2q_R$, while two halves that are independent given the design (disjoint repetitions, or disjoint items, with the same error probability) disagree in sign with probability $2q_{R/2}(1 - q_{R/2})$, so $\mathbb{E}[\mathrm{DIS}_0] = 2q_{R/2}(1 - q_{R/2})$. Hence $\mathbb{E}[\mathrm{DIS}_0] \geq \mathbb{E}[\mathrm{DIS}]$ whenever $q_{R/2}(1 - q_{R/2}) \geq q_R$. Under Gaussian noise with standardised gap $z = |\Delta|\sqrt{R}/\sigma$, so that $q_R = \Phi(-z)$ and $q_{R/2} = \Phi(-z/\sqrt{2})$, this holds exactly for $z \geq 0.843$, i.e. $q_R \leq 0.1996$ (the statement rounds inward to 0.199), and fails as $z \to 0$, where $q_R \to \frac{1}{2}$ and $\mathbb{E}[\mathrm{DIS}] \to 1$ while $\mathbb{E}[\mathrm{DIS}_0] \to \frac{1}{2}$: the crossing that Appendix C locates at a true gap near 0.04 in our geometry. The item floor is the same argument with I in place of R. □

Proposition 3. Exactness is the standard randomisation argument: under H_0 all design-label assignments within a (m, i) stratum are equally likely, so the permutation distribution of any statistic is its exact null distribution. Conservativeness: write $Y_{m,d,i,r} = \mu_{m,i} + \beta_d + \varepsilon$; β_d cancels in the observed $\Delta_d = b(m_1, d) - b(m_2, d)$, but the permuted cell means average draws across designs, so the permuted null variance includes $\mathrm{Var}_d(\beta_d)$, stochastically enlarging the reference distribution. □

Proposition 4. Fix $\Gamma \geq 1$ and minimise $w \mapsto \sum_d w_d \Delta_d$ over the box $[1, \Gamma]^S$ (scaling w leaves the sign unchanged, so the box is WLOG). The objective is linear and separable; the minimiser sets $w_d = \Gamma$ where $\Delta_d < 0$ and $w_d = 1$ otherwise, giving minimum $S^+ - \Gamma S^-$, which is ≤ 0 iff $\Gamma \geq S^+/S^-$; the infimum is attained there, so it is a minimum. If $S^- = 0$ the minimum is $S^+ > 0$ for every Γ, the feasible set is empty and $\Gamma^* = \infty$. □

Proposition 5. (1) For $S \subseteq S'$, $\{\text{L1 on } S'\} \subseteq \{\text{L1 on } S\}$ because the former is an intersection over a superset of per-design rejection events; monotonicity therefore holds for every joint distribution of the per-design statistics, with no independence assumption. If those statistics are independent the probability factorises to $\kappa(n)^k$, strictly decreasing and geometric. If they are positively orthant dependent, which is what sharing items across designs induces, $\Pr[\bigcap_d R_d] \geq \prod_d \Pr[R_d] = \kappa(n)^k$, so the geometric rate is a lower bound. (2) With designs drawn from a population, the two-stage variance identity gives $\mathrm{Var}(\bar{\Delta}) = (\sigma_\delta^2 + \sigma_\varepsilon^2/n)/k$ when every design draws fresh items; when the same I items serve every design, the item×model component σ_I^2/I is common to all designs and is not divided by k, which adds it as a floor. The power of the two-sided test on $\bar{\Delta}$ (the CI of §4.5) increases as the variance falls for any $\mu_\delta \neq 0$, up to that floor. (3) At $k = 1$ the two events coincide, and L1 implies L2 at every k, so $\Pr[\text{L1}] \leq \Pr[\text{L2}]$ throughout; as k grows $\Pr[\text{L1}]$ falls and $\Pr[\text{L2}]$ rises toward the value its item floor allows, so the curves diverge from their common $k = 1$ point. □

THE DESIGN-BUDGET DERIVATION

Now hold the total call budget $B = k \cdot n$ fixed, so that k designs at n observations each trade off against one another; per-observation noise is σ_ε^2. The distribution over δ_d is a device for computing power over *which designs an auditor happens to run*, what our sweep samples, not a claim that a deployment population exists.

Corollary 1 (allocation)**.** *Substituting $k = B/n$: (1) for L2,* $\mathrm{Var}(\bar{\Delta}) = (n\sigma_\delta^2 + \sigma_\varepsilon^2)/B$ *(plus the item floor of Prop. 5(2), which no allocation of B moves) is minimised at the smallest admissible n, the least depth at which a cell mean is defined, and the budget for fixed power does not grow with k:* maximal breadth, minimal depth. *(2) For L1,* $\Pr = \kappa(n)^{B/n}$ *is* increasing *in n, so the unconstrained optimum is $k = 1$: certification is maximised by shrinking the design set until L1 degenerates into L3 (Prop. 5(3)). The design set is therefore not a free parameter but is fixed by the invariance the auditor claims; holding k fixed and requiring* $\Pr[\text{L1}] \geq \gamma$ *gives $\kappa(n) \geq \gamma^{1/k}$, hence $n = \Theta(\log k)$ and $B = \Theta(k \log k)$. This is the homogeneous-gap case $\delta_d \equiv \mu_\delta$; under Prop. 5's heterogeneous δ_d the product is governed by* $\min_d \delta_d$*, so* $\Theta(k \log k)$ lower-*bounds the L1 budget and no n suffices once some $\delta_d \leq 0$.*

Corollary 1. (1) is the variance identity of Prop. 5(2) with $k = B/n$ substituted. (2) Write $f(n) = (B/n) \log \kappa(n)$. Since κ is increasing and $\kappa < 1$, both B/n and $|\log \kappa(n)|$ are decreasing in n, so $|f|$ is decreasing and f is increasing: the unconstrained maximiser is the boundary $n = B$, i.e. $k = 1$, at which L1 coincides with L3 by Prop. 5(3). Given k, $\kappa(n) \geq \gamma^{1/k} = 1 - |\log \gamma|/k + O(k^{-2})$; with per-design power $\kappa(n) = \Phi(\sqrt{n}\,\mu_\delta/\sigma - z_\alpha)$ this needs $\sqrt{n}\,\mu_\delta/\sigma - z_\alpha \geq z_{|\log\gamma|/k} \sim \sqrt{2 \log k}$, hence $n = \Theta(\log k)$ and $B = kn = \Theta(k \log k)$. □

C SYNTHETIC VALIDATION OF THE ESTIMATOR STACK

Before any API call, the full pipeline was run on synthetic panels with planted ground truth (six checks, C1–C6): (C1) permutation-test size under a null with large design main effects but no interaction: 0 rejections at $\alpha = 0.05$ across 40 null pairs (95% upper bound 0.072), with the DIS_0 floor above realised noise-only DIS (0.161 vs 0.133), confirming Prop. 2's conservative direction, and (C1b) 0.040 across 100 null pairs under the sharp exchangeable null with no design effect of any kind, so the test sits at nominal size where Prop. 3's variance inflation is inactive and below it where it is active; (C2) power 1.00 at interaction sd 0.4; (C3) zero false L1 certifications in 25 null pairs (95% upper bound 0.113); (C4) L2 bootstrap CI coverage 1.000 at nominal 0.95; (C5) Γ^* closed form vs. an independent bisection on Γ whose inner box-constrained minimisation is an exhaustive search over all 2^k weight vertices for $k \leq 12$ and the greedy vertex otherwise: max relative error 9.7×10^{-11} over 300 random contrast vectors of length 4–39; (C6) $\mathcal{N} \times \mathcal{F}$ variance shares and the exact orbit average recovered on noise-free cells.

We add the check those six leave open, L1's *power* at the configuration the paper actually runs (`code/power_and_intervals.py`). On synthetic panels with $k = 96$ designs, $R = 6$, and 16 items shared across all designs, so that per-design statistics are positively dependent exactly as in the real run, with per-design true contrasts $\delta + \mathcal{N}(0, 0.05^2)$, the L1 certification rate is 0.000 at $\delta = 0.05$ and 0.10, 0.070 at 0.20, **0.985** at 0.30, and **1.000** at 0.40 (200 trials each). At $\delta \leq 0.10$ that spread reverses at least one of the 96 designs in most trials (all at 0.05, about nine in ten at 0.10), so the zeros there are correct non-certifications rather than power; at $\delta \geq 0.20$ a reversed design occurs in under 1% of trials, so 0.070 at 0.20 is low power. With the 0.30 gap but one design reversed by construction the rate is 0.000, and with three reversed 0.000. The test is therefore well powered for genuinely separated pairs at this scale and collapses precisely when uniform dominance fails, which is what licenses reading a non-certification at a large gap as design-relativity rather than as low power, and what bounds how far the reading can be pushed at gaps below about 0.25. The floor's own scale was checked the same way: on pure-noise panels with no design effect, $\mathbb{E}[\mathrm{DIS}]$ vs DIS_0 is $0.94/0.50$, $0.63/0.46$, $0.50/0.43$, $0.25/0.33$, $0.07/0.18$ and $0.01/0.05$ at true gaps 0, 0.02, 0.03, 0.05, 0.08 and 0.12: the floor is anti-conservative below a gap of roughly 0.04 and conservative above it, which is why the triage pairs it with the permutation test (whose conjunction has a 1.9% false design-limited rate on the design-invariant null). We calibrate the interval-separation rule the same way. Declaring two models separated when their Imbens–Manski sets do not overlap fires for 0.003 of pairs whose identified sets exactly touch, 0.017 at a true separation of 0.05, 0.100 at 0.10, 0.827 at 0.20 and 1.000 at 0.30 (300 trials each, at the widths realised in our data): the rule is conservative at its nominal 5% and has real power only above a gap of roughly 0.2, so the 3–25% separation rate in §5.2 is a lower bound. The validation script and its frozen output ship with the artifact and are re-run by the reproduction command before any paid stage executes.

D CORPUS CONSTRUCTION AND PER-BENCHMARK DETAIL

Table 2: The published record, re-analysed against its own design variants. The *broad* family contains the strict one and adds tables whose crossed facet is a sub-task (a construct, not an elicitation design), so it bounds how far a ranking travels rather than evidencing fragility.

	strict	broad
tables / model pairs	17 / 612	28 / 1,287
pairs whose ordering reverses on $\geq$1 of the paper's own variants	**25.8%**	**54.3%**
median (within-model design range $\div$ cross-model spread)	0.50	0.95
tables where $>$1 model attains rank 1 under some design	70.6%	82.1%
tables carrying per-cell repetitions	0	0

The release carries five further CSVs from two sources whose numbers are published only as raster figures, which enter no analysis. Each source table was extracted twice (PDF text layer; word-coordinate reconstruction) and cross-checked cell-by-cell. Published marginals were recomputed from transcribed cells: 32/32 GAMA prompt-version marginals to within last-digit rounding (4 differ by 0.06, consistent with the source averaging unrounded values); SycEval's 58.19/43.52/14.66% headline reproduced, recovering the unpublished denominator $N = 15{,}345$; ELEPHANT's 11 means to ± 0.005; both Haase et al. (2026) decompositions to 100.00%. Three errors in source papers (a mislabelled column, a marginal off by a transposition, an inconsistent rounding) are documented in the corpus README and left uncorrected in the raw transcription with a correction layer applied separately. Median Γ^* among reversible pairs is 5.1 (strict) / 3.7 (broad); Condorcet cycles appear in 5.9%/14.3% of strict/broad tables and 6.2%/16.5% of pairs reverse under a $< 2\times$ re-weighting. Per-benchmark broad-family reversal rates: ELEPHANT 89%, CogMir 88%, In Agents We Trust 88%, GAMA 53%, Haase 43%, DarkBench 40%. The GAMA re-report under the proposed standard (§6): design means and intervals GPT-4o 68.6 [60.0, 74.9], Gemini-1.5-Pro 67.8 [64.4, 71.6], LLaMA-3.1-70B 65.0 [57.9, 69.8], GPT-3.5 46.5 [42.4, 52.4]; the three top-model pairs overlap pairwise. Exactly the three pairs involving GPT-3.5, the model that finishes last under all five designs, meet $\mathrm{L1}^{\mathrm{pt}}$ (the point-estimate condition; GAMA publishes no repetitions, so the tested L1 of §4.5 cannot be run on it); every comparison among the top three is L2-only, with Γ^* of 1.54, 3.25, and 4.81. The design-robust conclusions are the ones that did not need a benchmark.

E RESOLVING POWER OF THE PUBLISHED RECORD'S DESIGN COUNTS

The annotator floor under correlated errors. §3's floor assumes the three annotators mislabel independently. Repeating the simulation with a Gaussian copula giving their latent error scores equicorrelation ρ: the floor falls only from 0.108 ($\rho = 0$) through 0.107, 0.105, 0.103, 0.098 to 0.094 ($\rho = 0.7$). Correlated errors therefore make the record look slightly *more* design-limited, not less, and the §3 conclusion is not an artefact of the independence assumption. The floor is a mean over 546 pairs and is stable at this simulation size; the per-pair count is a tail statistic and is quoted only from the main analysis, which uses ten times the draws.

Panel composition and Monte-Carlo noise. Re-scoring the 144 legacy–legacy pairs inside the 16-model panel of §5.3 changes no DIS value at all (exactly, to machine precision) and moves two L1 and four L2 verdicts, so 138 of 144 are invariant to a change of panel composition that is irrelevant to the pair. Every moved L2 interval has an endpoint within 0.01 of zero, and re-certifying those six pairs at four seeds inside the 9-model panel alone moves three of them again: the residue is bootstrap noise at finite B, not a panel effect. A benchmark adopting the standard should fix its resampling seed per claim, so that the certificate is a deterministic function of that claim's data (`code/panel_composition_stability.py`).

Both floors, per domain. Mean DIS / $\mathrm{DIS}_0^{\mathrm{rep}}$ / $\mathrm{DIS}_0^{\mathrm{item}}$, then design-limited under the item floor (and under the repetition floor): `syco` 0.246/0.073/0.300, 7 (25) of 36; `darkpat` 0.410/0.114/0.500, 12 (31); `alloc5` 0.678/0.138/0.219, 36 (36); `alloc` 0.361/0.086/0.187, 26 (29); `dbreal` 0.483/0.067/0.205, 27 (35). Both floors are split-averaged Monte-Carlo estimates over a fixed collection: re-drawing the splits at four seeds moves a domain mean by at most 0.002 for $\mathrm{DIS}_0^{\mathrm{rep}}$ and 0.007 for $\mathrm{DIS}_0^{\mathrm{item}}$, so the domain ordering is unaffected; the values here are those Table 1 prints. The item floor exceeds the repetition floor everywhere, by $1.6\times$ in `alloc5` but $4.4\times$ in `darkpat`, and it exceeds mean DIS itself in `syco` and `darkpat`, the two domains where 16 items are too few to separate design from item sampling.

Triage by model gap. Design-limited share against $|\bar{\Delta}|$: 92% below 0.04 (37 of 40), 58% in $[0.04, 0.10)$ (28 of 48), 29% at or above 0.10 (16 of 56), and within that last bucket 14% at or above 0.20 (4 of 29). $\mathrm{DIS} = 0.662 - 2.088\,|\bar{\Delta}|$ ($R^2 = 0.54$). Design-limitation is concentrated where models are close, which is where leaderboards make their contested claims, and where the floor comparison of Prop. 2 is weakest, so the verdict there rests on the gap-restricted count of §5.1 (44 of 104 above 0.04).

Decoding settings. Every arm was requested at $T{=}1$ except two frontier models: `qwen-max` and DeepSeek-Reasoner were requested at $T{=}0$, and the reasoner endpoint documents that it ignores the parameter in thinking mode; o4-mini, GPT-5 and GPT-5-mini accept only $T{=}1$ and ran at their lowest reasoning effort. The completion cap is 16 tokens for the core panel, 24 for the Claude models and `qwen-max`, and 2,000 for the reasoning models, whose visible answer follows their reasoning. Parse rates (the share of responses the strict parser reads; unparsed responses leave the cell mean) exceed 0.96 for every model in every domain except these: DeepSeek-Reasoner reads at 0.89 overall (0.82 in `alloc`, 0.76 in `alloc5`, and as low as 0.55 under the business and casual framings), and in `alloc5` GPT-5-mini reads at 0.91 and qwen-turbo, GPT-4.1-nano and GPT-5 at 0.79, 0.81 and 0.86, the numeric-label collapse of §5.2. The frontier coefficient of §5.3 is therefore estimated with two of seven models at a different temperature and one with thinner cells; we report it as collected. $T{=}1$ is the setting that puts the most sampling noise into the repetition floor, while most audited benchmarks report near-deterministic runs. Replicating the two domains the item floor serves worst at $T{=}0$, everything else fixed (165,888 calls, $3.42), separates the two effects: the repetition floor falls by 39% and 45%, exactly as removing resampling should, while DIS does not fall at all.

domain	T	DIS	$\mathrm{DIS}_0^{\mathrm{rep}}$	$\mathrm{DIS}_0^{\mathrm{item}}$	cert	design-lim
`syco`	1.0	0.246	0.074	0.301	3	8
`syco`	0.0	0.255	0.045	0.331	2	7
`darkpat`	1.0	0.410	0.115	0.507	0	12
`darkpat`	0.0	0.448	0.063	0.484	0	15

So the instability the paper measures is a property of the models under a change of design, not of the decoder: greedy decoding removes the sampling component and leaves the design component where

it was. Both columns are single-seed splits, so the $T{=}1$ rows differ from Table 1's split-averaged floors by at most 0.007 and by one verdict in `syco`, and the movements in the verdict counts ($8 \to 7$, $12 \to 15$) are within the Monte-Carlo noise Appendix F quantifies for the certificates themselves.

Item-count curves. Nested prefixes of one collected bank per arm; 96 designs, 9 models, 36 pairs, $R = 4$ (`dbreal`) and $R = 6$ (`syco`) throughout, so only the item count varies. Table 1's `dbreal` row is this curve's $n = 96$ point. `syco` at $n = 16$ matches Table 1's configuration on a fresh collection and reproduces its verdicts exactly (7 of 36 design-limited); being a fresh bank, its floor estimate differs within replication noise (0.294 vs 0.300).

`dbreal` (DarkBench's own prompts)						`syco`					
items	$\mathrm{DIS}_0^{\mathrm{it}}$	DIS	cert	noise	design	items	$\mathrm{DIS}_0^{\mathrm{it}}$	DIS	cert	noise	design
12	0.500	0.526	0	16	20	16	0.294	0.247	3	26	7
24	0.391	0.554	0	8	28	32	0.215	0.223	5	20	11
48	0.266	0.528	0	8	28	48	0.173	0.194	7	16	13
96	0.205	0.483	0	9	27	96	0.129	0.199	10	10	16

The two arms differ in what the items buy. On DarkBench's material DIS is large and roughly flat, so items mainly lower the floor and the design verdict saturates once the floor is clear of the signal. In `syco` the signal is small and items sharpen both sides, moving claims out of *noise-limited* into both *certified* and *design-limited*: an item budget decides how much of a benchmark is decidable, not which way it decides.

The $\mathcal{N}/\mathcal{F}$ boundary. Which side of the boundary a generator sits on decides whether its variation is averaged away or reported as an interval, and the boundary is a normative judgement rather than a derived one. Promoting each nuisance generator into $\mathcal{F}$ in turn (12 variants over four domains, the design set untouched, so DIS, the permutation test and L1 cannot move) leaves the $\mathcal{N}$-symmetrised point estimate *exactly* unchanged — a grand mean over a fully crossed grid does not depend on how its axes are partitioned — while mean interval width rises by up to $\times 2.65$ and the number of interval-separated pairs falls in *all* 12 variants (`syco` 17 to 8–14, `darkpat` 16 to 5–13, `alloc5` 11 to 0–8, `alloc` 16 to 4–8). A stricter reading of which facets a competent agent must be indifferent to therefore makes the paper's conclusions more conservative, not less, so the headline cannot be an artefact of a self-serving split.

Triage calibration. On the design-invariant null (96 pseudo-designs resampled from one real design's repetitions) the whole triage returns design-limited 1.9% of the time, 11 of 576.

Item-count saturation. DarkBench's item curve plateaus by 24–48 items; `syco`'s is still improving at 96, so our own 16 were too few to decide most of it. **The `alloc5` gap exception.** `alloc5` keeps $\mathrm{DIS} = 0.53$ over its 5 pairs beyond 0.10, which is the widest cut its 0.157-wide panel admits, and the label-compliance failure of §5.2 does not explain it: over the 64 designs that avoid that alphabet DIS rises to 0.701 with $\mathrm{L1} = 0$.

Nuisance versus framing. Nuisance alone (framing fixed) yields mean DIS of 0.18–0.61, exceeding framing-only instability in two domains of four: transformations to which a competent agent must be indifferent move cross-model conclusions about as much as personas do. **Partial arms.** Gemini 2.5 Flash reached 209 of 36,864 cells and an earlier Google pilot fewer; both ship, enter no analysis, and are excluded by name in the reproduction command.

Frontier models, per domain. In `syco` the frontier models are not separable from one another at all; in `alloc` they are more stable, lifting its L1 rate from 3/36 on the 9-model panel to 23/120 on the 16-model one, the one place a prediction fixed before the run failed (prediction log in the artifact).

Frontier models, gap-adjusted. Frontier–frontier pairs are closer than legacy–legacy pairs (median $|\bar{\Delta}|$ 0.028 vs 0.075), so their higher raw DIS (0.484 vs 0.424) is confounded. Regressing DIS on $|\bar{\Delta}|$ and a frontier indicator over the 228 same-class pairs (84 frontier–frontier, 144 legacy–legacy) gives a frontier coefficient of $+0.018$ (se 0.029, $t = 0.62$, 95% CI $[-0.038, +0.074]$). Pairs sharing a model, and one dyad measured in four domains, are dependent, so we also aggregate each dyad over domains ($+0.023$, se 0.026, 57 dyads) and bootstrap by *model*, resampling the seven frontier and nine legacy models and keeping induced dyads: se 0.050, 95% CI $[-0.072, +0.124]$. The point estimate is stable across all three; the abstract quotes the dyad estimate with its model-clustered se. A first pass with three frontier models and 12 such pairs gave $+0.070$ (se 0.063): same sign, but an

interval too wide to bound anything. We claim that design-relativity is not a small-model artefact, and that any attenuation with capability is smaller than 0.072 in DIS under the model-clustered bootstrap. **Resolution at the record's design counts.** Fraction of the 36 pairs per domain called *design-limited* against the repetition floor $\mathrm{DIS}_0^{\mathrm{rep}}$ (the floor the published record can supply, with the item set fixed at 16), when that triage is run on random subsets of k of our 96 designs at R repetitions per cell; at $k = 96$, $R = 6$ the rates match the parenthesised counts of "Both floors, per domain" above, not Table 1's dual-floor column. Each interior cell averages 8 random design subsets; $k = 96$ is the full panel. The published record operates at $k = 3$ (DarkBench's annotators), $k = 5$ (GAMA's paraphrases) or $k = 6$ (GAMA's temperatures), with $R = 1$ in every crossed table. The rate is monotone increasing in k in all four domains at $R = 6$, and in 94 of the 96 adjacent (k, k') steps across all four depths; the two exceptions are both k=5 $\to$ 6 (-0.049 in `syco` at R=3, -0.007 in `alloc` at R=4), within the draw noise of 8 random subsets. What the record can detect is bounded by how many designs it prints, so its *noise-limited* verdicts carry no information about whether a design effect exists. The R axis moves the rate far less than k does, and not monotonically at $k = 96$, where a single subset of 36 pairs is evaluated and a 0.03 difference is one pair.

domain	R	k=3	k=5	k=6	k=12	k=24	k=48	k=96
syco	2	0.19	0.23	0.27	0.38	0.51	0.54	0.58
	3	0.23	0.35	0.30	0.43	0.50	0.53	0.53
	4	0.21	0.34	0.37	0.46	0.54	0.62	0.69
	6	0.21	0.33	0.44	0.48	0.55	0.64	0.69
darkpat	2	0.22	0.25	0.32	0.58	0.70	0.82	0.89
	3	0.26	0.48	0.49	0.68	0.76	0.79	0.81
	4	0.42	0.50	0.52	0.70	0.76	0.82	0.86
	6	0.41	0.56	0.59	0.72	0.82	0.85	0.86
alloc5	2	0.38	0.75	0.76	0.93	0.96	0.99	1.00
	3	0.53	0.76	0.82	0.92	0.99	0.99	1.00
	4	0.59	0.81	0.86	0.94	0.99	1.00	1.00
	6	0.62	0.77	0.90	0.98	1.00	1.00	1.00
alloc	2	0.27	0.52	0.52	0.66	0.78	0.83	0.86
	3	0.33	0.39	0.48	0.66	0.75	0.82	0.83
	4	0.28	0.51	0.50	0.62	0.76	0.81	0.86
	6	0.32	0.43	0.49	0.71	0.74	0.80	0.81

F FULL FORMAL-RUN TABLES

Nuisance-only versus framing-only instability, mean DIS (the 12 nuisance levels at the neutral framing; the 8 framings at one fixed nuisance level): `syco` 0.183 vs 0.149, `alloc5` 0.611 vs 0.455, `darkpat` 0.340 vs 0.392, `alloc` 0.241 vs 0.295. A replicate-only null of 96 pseudo-designs gives 5.7–12.3 distinct orderings (p95 12–31), which the observed 73–94 exceeds in every domain. Permissive-parser sensitivity: one L2 certificate and one sign-of-mean differ across the 144 claims, and mean DIS moves by ≤ 0.006. **Resampling settings.** Table 1 uses $B = 2000$ bootstrap resamples for each per-design test and for the L2 interval, 60 half-splits per floor and 400 permutations per pair (seed 20260811); the Bonferroni statement of §5.1 and the gap analysis of Appendix E use 6,000 permutations (seed 20260822), as do the item curves (seed 20260827) and the frontier triage ($B = 1000$, seed 20260828). Every arm has an even R and an even item count, so the two halves of a split are equal in size. DIS splits an exact tie between p^+ and p^-; the floors count a half whose contrast is exactly zero as disagreeing with the other half, so a tie raises the floor rather than being split, which is the conservative direction for the triage. Tables 3–5 report, per domain: point estimates with identified intervals and single-design (L3) ranges per model; the facet decomposition and both parser arms; and compliance by nuisance generator in `alloc5`, including the numeric-label collapse of §5.2. **Denominators.** Two families of statistic have different denominators in `alloc5`, and we state both. Pair-level quantities (DIS, DIS_0, the permutation test, L1/L2, Γ^*) are computed over the designs at which both members of a pair are defined, so all 9 models and all 36 pairs enter in every domain (144 in total). Estimand-level quantities (point estimate, $\mathcal{I}(m)$, interval separation) require the *complete* $\mathcal{N}$-orbit, since an orbit average over part of $\mathcal{N}$ is not the estimand of Def. 1; Qwen-turbo has 2 of 96 `alloc5` cells with no parseable response, so `alloc5` contributes 8 models and 28 pairs to those. This is the sense in which it "drops out" in §5.2: not removed from the panel,

but possessing no estimand to compare. Interval separation, plug-in: `syco` 17/36, `darkpat` 16/36, `alloc5` 11/28, `alloc` 16/36; with Imbens–Manski confidence sets: 1/36, 3/36, 7/28, 5/36.

Leaderboard counts for `alloc5` in Table 1 use the 94 designs at which every model's cell mean is defined. Over all 96 they would read 92 distinct orderings rather than 90, because `argsort` places an undefined value last and so silently ranks Qwen-turbo bottom at the 2 designs where it gave no parseable answer, an artifact of the missing value rather than a measurement. The rank-1 count is unaffected. **Budget-curve endpoints.** At $R = 3$, L1 runs $k = 2 \to 36$: 0.37→0.06 (`alloc`), 0.33→0.10 (`syco`), 0.13→0.00 (`alloc5`), 0.07→0.00 (`darkpat`); L2 rises 0.41→0.76, 0.38→0.68, 0.16→0.63, 0.09→0.37. At $R = 6$, comparable with Table 1's $k = 96$, L1 at $k = 2 \to 36 \to 96$ is 0.34→0.12→0.08 (`syco`), 0.44→0.07→0.08 (`alloc`), 0.26→0.00→0.00 (`alloc5`), 0.12→0.00→0.00 (`darkpat`), monotone except `alloc`'s last step (0.07 vs 0.08), within the spread of a 12-draw estimate. Curves use 12 random design draws per point and 200 bootstrap resamples. In the plotted curves (Figure 3, $R \leq 3$), 158 of the 168 adjacent steps are monotone; the ten exceptions are all ≤ 0.030 (the largest is `alloc5`'s L2 at $k = 4 \to 6$, $R = 3$), seven of them in `alloc5` where L1 is at zero, and at the deepest plotted $R = 3$ the L1 curve is monotone in three domains with one up-step of 0.012 in `alloc5`. These are artifacts of drawing non-nested subsets, which Prop. 5(1) does not cover.

Tie conventions and $\mathcal{N}$-set sensitivity. Mean DIS with ties split vs. dropped: `syco` 0.246/0.237, `darkpat` 0.410/0.302, `alloc5` 0.678/0.628, `alloc` 0.361/0.353; the gap tracks how often ties occur (33 of 36 `darkpat` pairs have a tied design, 15.1 on average, against 8 of 36 and 1.1 in `alloc`). Dropping one level of one nuisance generator (7 variants per domain) shifts a point estimate by at most 0.098 (`syco`), 0.057 (`darkpat`), 0.158 (`alloc5`) and 0.030 (`alloc`), with means of 0.008–0.020, and changes the induced ordering in 4, 4, 6 and 4 of the variants. Certification under a design-invariant null (96 pseudo-designs resampled from one real design's repetitions): L1 0.41/0.04/0.50/0.51 and L2 0.46/0.08/0.62/0.66 for `syco`/`darkpat`/`alloc5`/`alloc`. `alloc5` over its 64 non-numeric-label designs: DIS $= 0.701$, 61 of 64 distinct orderings, 6 rank-1 models, L1 $= 0$.

Abstract-level claim strength. Coding each audited paper's cross-model claim: DarkBench "some LLMs are explicitly designed to favor their developers' products" (unconditional); SycEval "Gemini exhibiting the highest rate" (explicit ordering); ELEPHANT "LLMs consistently exhibit high rates of social sycophancy" (unconditional level claim over 11 models); ImpossibleBench, a capability trend in cheating rate; GAMA, a ranked leaderboard over eight games; CogMir and In Agents We Trust, cross-model level comparisons. Seven of eight therefore assert something only L1 licenses. The exception is Haase et al. (2026), which reports variance shares and asserts no ranking. Every quoted fragment was matched against the source PDF text.

Compliance-conditioned estimands. Recomputing every `alloc5` estimand with the numeric-label designs removed (the alphabet whose compliance collapses) shifts the point estimates by at most 0.019 and on average 0.008; the largest movers are doubao-lite $0.587 \to 0.606$, doubao-mini $0.541 \to 0.555$ and gpt-3.5-turbo $0.657 \to 0.645$.

Infinite Γ^*: strict or tied. Per domain, (strict, tied, finite) counts of the 36 pairs are `syco` (9, 1, 26), `alloc` (5, 0, 31), `darkpat` (1, 8, 27), `alloc5` (0, 1, 35); medians over the finite values are 21.6, 9.45, 9.23, 4.32. In `darkpat` almost every infinite tilt is a tie rather than dominance, because with 16 items $\times$ 6 repetitions the per-design cell mean lies on a coarse lattice and near-zero-rate pairs are exactly equal at many designs, which is why §4.4 reports the finite median separately.

The published-record triage in full. GAMA reports five repeated runs (Tables 4/10/14/18) and five paraphrases (Tables 6/12/16/20) over the same eight games and four models. GAMA publishes only whole-run aggregates, so its floor is the index of Def. 3 applied across the five repeated runs (values in multiples of 0.4), not the split-half estimator DIS_0 of §4.3; we write DIS_0 for it because it plays the same role. Per-pair DIS across paraphrases vs. across repeated runs: GPT-3.5/GPT-4o 0.00 vs 0.00, GPT-3.5/Gemini 0.00 vs 0.00, GPT-3.5/LLaMA 0.00 vs 0.00, GPT-4o/Gemini 0.40 vs 0.80, GPT-4o/LLaMA 0.80 vs 0.80, Gemini/LLaMA 0.40 vs 0.40. No pair exceeds its floor, and the three pairs that do move are exactly the three top-model pairs whose Γ^* is 1.54, 3.25 and 4.81. Machine-readable versions of every table are generated by one command from the raw data.

Table 3: Point estimate, identified interval, and single-design range per model (strict parser). The point estimate pools the exact $\mathcal{N}$-orbit averages over $\mathcal{F}$; the interval is $[\min_f, \max_f]$ of the orbit curve; the L3 range is the spread of the 96 single-design values a conventional benchmark could have reported.

	syco			darkpat		
model	point	interval	L3 rng	point	interval	L3 rng
deepseek-chat	0.278	[0.225, 0.339]	0.552	0.058	[0.022, 0.122]	0.271
doubao-lite	0.088	[0.027, 0.317]	0.552	0.078	[0.013, 0.112]	0.165
doubao-mini	0.141	[0.082, 0.193]	0.375	0.113	[0.067, 0.170]	0.228
gpt-3.5-turbo	0.562	[0.441, 0.670]	0.542	0.181	[0.157, 0.226]	0.354
gpt-4.1-mini	0.033	[0.005, 0.076]	0.156	0.027	[0.011, 0.049]	0.177
gpt-4.1-nano	0.317	[0.275, 0.388]	0.583	0.009	[0.001, 0.019]	0.104
gpt-4o-mini	0.109	[0.031, 0.350]	0.417	0.040	[0.000, 0.078]	0.146
qwen-plus	0.042	[0.011, 0.181]	0.260	0.000	[0.000, 0.000]	0.000
qwen-turbo	0.313	[0.266, 0.470]	0.490	0.028	[0.005, 0.072]	0.125
	alloc5			alloc		
deepseek-chat	0.597	[0.539, 0.657]	0.971	0.841	[0.667, 0.951]	0.422
doubao-lite	0.587	[0.500, 0.755]	0.464	0.632	[0.537, 0.756]	0.440
doubao-mini	0.541	[0.506, 0.593]	0.341	0.586	[0.537, 0.665]	0.428
gpt-3.5-turbo	0.657	[0.607, 0.735]	0.818	0.511	[0.484, 0.532]	0.117
gpt-4.1-mini	0.505	[0.502, 0.509]	0.063	0.512	[0.500, 0.536]	0.139
gpt-4.1-nano	0.530	[0.513, 0.556]	0.323	0.523	[0.494, 0.610]	0.444
gpt-4o-mini	0.519	[0.495, 0.553]	0.187	0.501	[0.500, 0.506]	0.031
qwen-plus	0.500	[0.500, 0.501]	0.016	0.747	[0.532, 0.929]	0.498
qwen-turbo	—	—	—	0.567	[0.534, 0.608]	0.339

Table 4: Facet decomposition and parser sensitivity. DIS(N)/DIS(F): mean disagreement with only the nuisance (framing fixed) / only the framing (nuisance fixed) varying; removable: mean share of the per-model elicitation range deleted by exact $\mathcal{N}$-symmetrisation; strict vs. permissive parser arms.

domain	DIS(N)	DIS(F)	removable	$\text{parse}_{\text{strict}}$	$\text{parse}_{\text{perm}}$	$\text{DIS}_{\text{strict}}$	DIS_{perm}	$\text{L2}_{\text{strict}}$/$\text{L2}_{\text{perm}}$
syco	0.183	0.149	56%	0.997	0.997	0.246	0.246	69%/69%
darkpat	0.340	0.392	62%	0.996	0.996	0.410	0.410	36%/36%
alloc5	0.611	0.455	78%	0.955	1.000	0.678	0.672	78%/81%
alloc	0.241	0.295	60%	1.000	1.000	0.361	0.361	83%/83%

G THE ZERO-COST OPEN-WEIGHTS ARM

An additional arm reads the *exact* option distribution of every cell from open-weight model logits in a single forward pass (CPU-only, $0; six 135M–1.1B models, two domains, a reduced 36-design grid (6 nuisance $\times$ 6 framings) with 8 items). The models are the instruction-tuned Hugging Face checkpoints `HuggingFaceTB/SmolLM2-135M-Instruct`, `SmolLM2-360M-Instruct`, `SmolLM-360M-Instruct`, `Qwen/Qwen2.5-0.5B-Instruct`, `Qwen2-0.5B-Instruct` and `TinyLlama/TinyLlama-1.1B-Chat-v1.0`, run in float32; the domains are `syco` and `alloc5`, the two forced-choice domains whose answer is a single option token, so the cell's behaviour is the softmax over the option labels with no decoding. The finite-R arm draws $R = 6$ responses per cell as multinomial samples from that exact distribution, which is what a sampled API arm at temperature 1 would produce; the `meta` field of every released record carries the exact probabilities alongside the draw. Because cell means are then noise-free, DIS_{true}, the design instability at infinite repetitions, is available as ground truth, which the API arm cannot supply. Findings: the excess $\text{DIS} - \text{DIS}_0$ is a conservative lower bound on DIS_{true} for 97% of pairs, confirming Prop. 2's direction on real models; at $R = 6$, $|\text{DIS} - \text{DIS}_{\text{true}}| = 0.070$ on average while the floor is 0.168, confirming that DIS_0 must be used as a one-sided threshold, not subtracted as a "noise share"; Γ^* computed from sampled vs. noise-free cell means differs by a median $|\log \text{ratio}|$ of 0.067; and the permutation test rejects for 100% of pairs, all of which have nonzero true design variance (a power check; size was established synthetically). This arm also demonstrates that the framework applies without API access.

Table 5: Label-space compliance by nuisance generator, `alloc5` (share of responses that are a valid label in the alphabet the prompt used, averaged over the other generators). The numeric-label column is the collapse of §5.2.

model	lab=alpha	lab=lower	lab=numeric	opt=rev	opt=std	lay=block	lay=inline
deepseek-chat	1.000	1.000	0.998	0.999	0.999	0.999	0.999
doubao-lite	1.000	1.000	1.000	1.000	1.000	1.000	1.000
doubao-mini	1.000	1.000	1.000	1.000	1.000	1.000	1.000
gpt-3.5-turbo	1.000	1.000	1.000	1.000	1.000	1.000	1.000
gpt-4.1-mini	1.000	1.000	1.000	1.000	1.000	1.000	1.000
gpt-4.1-nano	0.998	1.000	0.438	0.809	0.814	0.784	0.840
gpt-4o-mini	1.000	1.000	1.000	1.000	1.000	1.000	1.000
qwen-plus	1.000	1.000	1.000	1.000	1.000	1.000	1.000
qwen-turbo	1.000	1.000	0.362	0.788	0.787	0.793	0.782

H COST ACCOUNTING

Model strings are recorded exactly as the serving API exposes them, a provenance caveat this paper's own argument makes salient. Core audit: 331,776 calls, 38.5M prompt tokens, 388K completion tokens, $8.544, rebuilt from per-request usage records (per-model breakdown in the artifact); the pilot's cost forecast erred by 1.2%. Frontier extension: 258,257 calls (258,048 on the $7\times4\times96\times16\times6$ grid plus 209 responses from a partial Gemini pilot that ships but enters no analysis), $187.674 over seven models, of which Claude Opus 4.7 alone is $97.60; the enlarged panel buys the interval of §5.3 and nothing else, and the three-model first pass ($25.16 of the $187.674) supports the same sign. DarkBench-item extension: 331,775 calls (one cell of the 331,776 returned no response), $17.403. Sycophancy item extension: 497,664 calls, $9.910. Decoding-temperature replicate: 165,888 calls, $3.417. Open-generation plus three-annotator extension: 13,808 generations ($1.792) over 16 designs, the crossing of a 4-level nuisance set (an opening greeting and a closing courtesy, each present or absent, the only surface material an open prompt admits) with four of the eight framings (neutral, assistant, terse, casual), and 47,142 verdicts ($25.825) from the annotators `claude-sonnet-4-6`, `gpt-4o-mini` and `qwen-max` at temperature 0 with a one-word YES/NO rubric per category and the option order randomised per call, of which 41,424 form the analysed rectangle and 5,718 come from two further judge candidates run on part of the grid ($2.66 of the $25.825), which ship but enter no analysis. With the pilots, the partial 48-item generation set and the earlier minimal-validation arm ($0.928 together), the entire project cost $255.49 of API spend and no GPU-hours. Every figure here except the judging pass is recomputed from the token counts carried by the shipped records themselves (`code/cost_from_records.py`), so it is a function of the artifact rather than of how collection happened to be scheduled; the judge pass stores verdicts without token counts, so its $25.825 is carried from the collection ledger. The core audit alone carries every claim in Table 1 and Appendix F except the `dbreal` row. The `dbreal` continuations were written by `claude-sonnet-4-6` at seed 20260821 and frozen; on the 24-item seed bank, `gpt-5-mini` and `qwen3.6-plus`, neither of which generated the items or sits on the nine-model `dbreal` panel, were each asked which continuation exhibits the named pattern with the option order randomised, and both identified the dark continuation in 23 of 24 items (the 24th is kept and flagged in the released bank). The 72 items added to reach 96 were generated by the same frozen procedure and are released with the same fields.

I THE CAPABILITY-TREND DISAGREEMENT TABLE

Seven papers make mutually inconsistent claims about whether design sensitivity falls with capability: two report a rise, none a clean fall, and five report no monotone trend, a within-paper sign flip, or no improvement with scale. Our per-model data (nine models, three within-family ladders) shows no monotone pattern in either direction across domains, and we decline the claim in both directions. Table 6 names the seven papers with their per-paper operationalisations, classified by this paper's taxonomy (N: nuisance; F: framing; X: neither semantics- nor information-preserving, such as few-shot count or chain-of-thought); the quoted fragments and their sources are given in full in `docs/DISAGREEMENT_TABLE.md` in the artifact. Call a reported trend *clean* when the paper varies one perturbation family only. Exactly one clean rise is reported (framing alone); the one

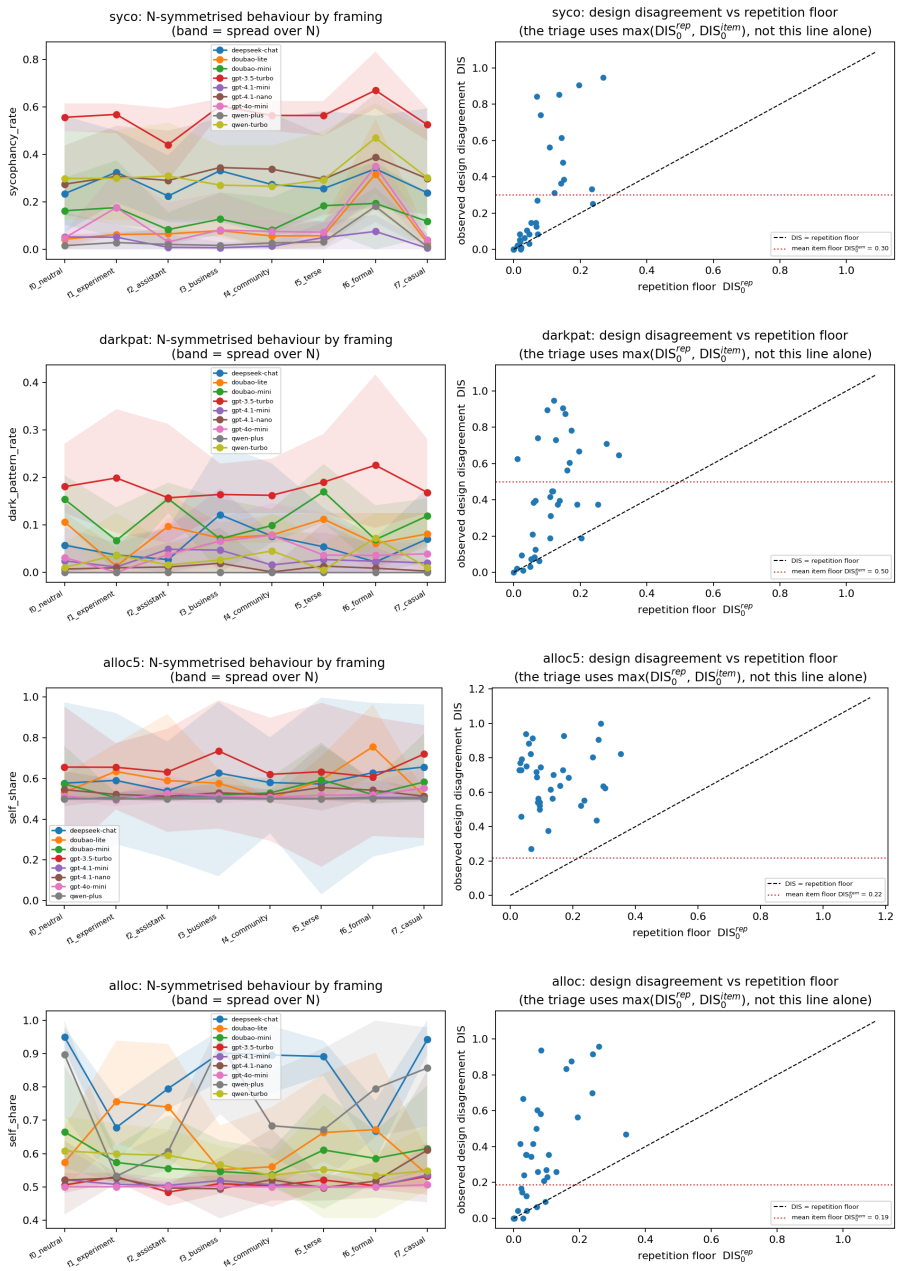

Figure 2: All four domains, top to bottom `syco`, `darkpat`, `alloc5`, `alloc`. **Left**: the estimand of Def. 1, per-framing exact $\mathcal{N}$-orbit averages (points) with the raw spread over $\mathcal{N}$ (bands); the identified interval is the range of the point curve and the band is what symmetrisation removes. **Right**: design disagreement DIS against the *repetition* floor $\mathrm{DIS}_0^{\mathrm{rep}}$ for all 36 pairs. Most sit above the diagonal (69–100% by domain) and the permutation test rejects exchangeability for every pair, but the triage of §4.3 uses $\max(\mathrm{DIS}_0^{\mathrm{rep}}, \mathrm{DIS}_0^{\mathrm{item}})$, under which 57 of 144 are noise-limited (§5.1); this panel therefore shows the weaker of the two floors, not the operative one.

candidate for a clean fall varies nuisance alone and flips sign across formats within its own family; no two papers vary the same family; and six of the seven measure no self-consistency floor, so none of the trends is normalised against run-to-run instability. That missing floor, not any N/F pattern, is the common defect.

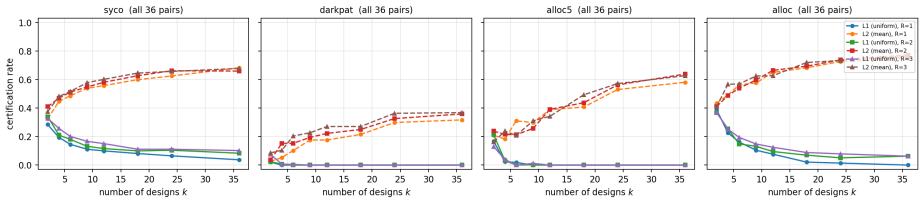

Figure 3: Budget curves for all four domains, computed over **every** one of the 36 model pairs per domain ($k \in \{2, 4, 6, 9, 12, 18, 24, 36\}$, $R \in \{1, 2, 3\}$; 12 random design draws per point, 200 bootstrap resamples). Each curve holds depth per design fixed at R repetitions and grows the design set k; total calls therefore grow with k, so this is the fixed-n monotonicity of Prop. 5, not the fixed-B allocation of Cor. 1. L1 (solid) falls in k and L2 (dashed) rises in every domain at every depth, up to ten adjacent steps of at most 0.030 out of 168, seven of them in `alloc5` where L1 sits at zero (Appendix F).

Table 6: Seven papers on whether design sensitivity falls with capability, classified by this paper's taxonomy. No two vary the same perturbation family and six measure no floor.

paper	models	perturbations varied	reported capability trend	floor?
Allen & Peterson (2026)	14 frontier API models	F only (three goal framings), one domain relabel	rises: newer models less robust	partial (reps, no floor statistic)
de Zarzà et al. (2026)	7 open models (Hermes-4, Qwen3, DeepSeek-R1, gpt-oss)	mixed: N (fact order), F (paraphrase, context), X (expand, contract)	rises with scale (inverse scale–robustness)	defined, not estimated
Bellibatlu et al. (2026)	13 judge models	F (instruction paraphrase) with an N control (A/B swap)	no monotone trend; within-family inversions	no (T=0, one run)
Karmakar & Chatterjee (2026)	15 open models, 1–8B	mixed: N (reorder), F (paraphrase, framing), X (few-shot)	no reliable trend, $\rho \in [-0.24, 0.47]$ across benchmarks	no (greedy, one run)
Cho et al. (2026)	Qwen2.5 0.5B–72B and others	N only (format, choice component, instruction, few-shot)	sign flips within the paper across formats	no (log-likelihood scoring)
Sclar et al. (2024)	LLaMA-2 7–70B, Falcon, GPT-3.5	N only (separators, casing, spacing)	no improvement with scale	no
Zheng et al. (2024)	LLaMA and LLaMA-2 families, Vicuna, Falcon, GPT-3.5	N only (option position and ID)	no consistent within-family pattern	no