

Moral Competence Before Moral Content: Why LLM Agents Lack the Prerequisites for Coherent Alignment

Arno Libert, Derck W.E. Prinzhorn and Daan R. Henselmans

Aithos Research Foundation, Keizersgracht 62, Amsterdam, 1015 CS,
the Netherlands.

*Corresponding author(s). E-mail(s): arno@aithos.org; daan@aithos.org;
Contributing authors: derck@aithos.org;

Abstract

AI alignment requires AI systems to adhere to human norms, values, or intentions. Under value pluralism there is no correct target, but a shared prerequisite is that the system’s behavior expresses a coherent *policy*: a mapping from situations to verdicts that is invariant while a situation’s morally relevant features are preserved, and sensitive when they change. We introduce four structural conditions for such coherent policies: verdict stability, monotonicity, decisiveness, and Pareto viability. Together they measure a form of moral competence that is evaluable from behavior alone, without reference to a moral standard or expert baseline, forming a structural floor for alignment rather than a normative target. We demonstrate the methodology on three simulated deployments featuring LLM-based agents facing moral dilemmas. Evaluating nine frontier models under a factorial design of five paraphrases, five escalation levels, and three dominance conditions, we show no model expresses a coherent policy across the three deployments: surface-form perturbation alone produces verdict-rate shifts of up to **99** percentage points at a single escalation level, and a model’s success on one scenario does not predict its competence on another. This suggests LLM-based agents are not currently the kind of object to which alignment can meaningfully apply.

Keywords: AI alignment; AI ethics; moral competence; value pluralism; LLM evaluation; agentic AI

1 Introduction

AI alignment asks how to make AI systems adhere to human norms, values, or intentions so they genuinely serve human interests (Russell, 2019). Since humans hold divergent values and commitments, which target is appropriate remains both technically and normatively open (Gabriel, 2020), and proposed answers span various frameworks: social-choice approaches aggregate stakeholder preferences into an ordering the system enacts (Conitzer et al., 2024; Ge et al., 2024); utility-based approaches treat the system as maximizing a coherent, ideally Pareto-optimal utility function (Desai, Critch, & Russell, 2018; Mazeika et al., 2025); fair-process approaches derive principles the system must apply (Fazelpour & Fleisher, 2025; Gabriel & Keeling, 2025; Huang et al., 2024); contractualist approaches target context-specific norms (Levine et al., 2026; Zhi-Xuan, Carroll, Franklin, & Ashton, 2025); and pluralistic approaches aim toward representation of human plurality (Kasirzadeh, 2024; Sorensen et al., 2024).

Across these proposals, one structural prerequisite runs prior: adhering to the target requires that the system’s behavior expresses a certain *policy*: a function from situations to verdicts. A *verdict*, on this reading, is the outcome selected by an AI system when it carries out a decisive action, while an *alignment target* is a specification of which policy the system should express, whether that is based on values, intents, principles, or norms.

An AI system given decisiveness over a set of outcomes always expresses a policy through its behavior. Since any way of handling a situation leads to *some* outcome, a system’s policy constitutes a mapping of situations to a verdict *distribution* at minimum. However, alignment targets, at least ones that normatively follow principles, intentions, norms, or values, also require the policy to be *coherent*: invariant across contexts that preserve all features of a situation relevant to the target, but sensitive to changes that the target deems decisive. Without a coherent policy, empirical observation of samples drawn from its verdict distribution—a system that is observed following instructions, expressing certain values, or exhibiting specific behavior—may not be representative of its performance across deployments, as the system’s verdicts may invert under meaningless variation or fail to adapt to morally relevant context.

This matters because LLMs are both notoriously unpredictable and sensitive to semantically empty prompt variation. Verdicts on moral dilemmas are typically inconsistent across samples (Krügel, Ostermaier, & Uhl, 2023; Scherrer, Shi, Feder, & Blei, 2023). Surface-level perturbation introduces further instability, both in moral dilemmas (Oh & Demberg, 2025) and capability tests (Sclar, Choi, Tsvetkov, & Suhr, 2024). Verdicts robust to lexical noise still flip under content-preserving point-of-view shifts and evaluation-protocol changes (van Nuenen & Sachdeva, 2026). Query phrasing has a larger impact on AI status disclosure rate than the model used, language, or scenario context (Gausen, Wallbridge, O’Dell, Summerfield, & Kirk, 2026) and sycophantic accommodation shifts verdicts that should remain stable (Sharma et al., 2024). Benchmark scores measured at a single configuration are correspondingly poor proxies for deployed behavior (Bean et al., 2025; Bengio et al., 2026).

While these trends are concerning when considering general-purpose LLM assistants, especially given human willingness to accept influence on moral decision-making (Dillion, Mondal, Tandon, & Gray, 2025; Krügel et al., 2023; Sharma, McCain,

(Douglas, & Duvenaud, 2026), agentic LLM deployment is particularly vulnerable. Paraphrasing an agentic scenario can swing the likelihood of engaging in illegal insider trading by up to 88.7 percentage points (see Appendix A.2), changing tool names to sound more benign significantly increases model propensity to engage in dangerous behaviors like self-proliferation and chemical weapon synthesis (Sehwag et al., 2025), and referring to an agent as “the AI system” instead of naming it has a larger effect on its propensity to engage in blackmail than removing its explicit goal conflict (Lynch et al., 2025).

This research introduces formal measurements to determine whether frontier LLM agents express coherent policies under realistic stochastic noise. We propose four structural conditions—verdict stability, monotonicity, decisiveness, and Pareto viability (Section 3.1)—that any system expressing a coherent policy must satisfy, regardless of the target it serves, and analyze three single-choice agentic dilemmas—without prescribing a correct option—that can be used to evaluate these properties.

The remainder of the paper is organized as follows. Section 2 situates the proposed metrics within the literature on pluralistic alignment and moral competence in LLMs. Section 3 specifies a factorial design to measure the four properties simultaneously across novel agentic scenarios and reports results across nine frontier models. Section 4 discusses implications for evaluation methodology and alignment architecture.

2 Background

2.1 Alignment under Value Pluralism

Contemporary alignment research concentrates around a small set of normative targets. Leike et al. (2018) propose AI systems should behave in accordance with the intention of their operators. Askill et al. (2021) build on this to formulate the helpful-honest-harmless (HHH) framework that has become a de facto standard. Subsequent work on Constitutional AI (Bai et al., 2022) and reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022) operationalizes refusal of dangerous requests and policy compliance as primary training objectives.

A growing literature observes that this framing is in tension with the context-dependent and fundamentally contested conditions alignment requires in practice (Conitzer et al., 2024; Dobbe, Krendl Gilbert, & Mintz, 2021; Fazelpour & Fleisher, 2025; Gabriel et al., 2024; Rudschies, Schneider, & Simon, 2020). Real deployment contexts are pluralistic, meaning they feature diverse, sometimes irreconcilable perspectives. This plurality is a feature of the moral landscape some philosophers hold to be a precondition of ethics itself (Arendt, 1958; Berlin, 1969). Aggregating numerous disagreeing values into one optimal ordering has well-known limits in social choice (Sen, 1970). These extend to alignment methods like RLHF, which faces several fundamental limits (Casper et al., 2023), implicitly aggregates over hidden context through a Borda rule (Siththaranjan, Laidlaw, & Hadfield-Menell, 2023), and must necessarily override personal preferences (Mishra, 2023). Setting alignment targets through democratic representation faces Arrow-style impossibilities of its own (Qiu, 2024), and—we argue elsewhere—fair multi-stakeholder alignment in a single system is impossible as such (Henselmans, Libert, Pier van Gosliga, & Zwart, 2026).

Recent work on pluralistic alignment removes the need for a unified target: rather than aggregating, different stakeholders may set their own. [Sorensen et al. \(2024\)](#) distinguish Overton, steerable, and distributional pluralism as alignment targets. [Kasirzadeh \(2024\)](#) defines a normative infrastructure that extends beyond value choices to include who has the authority to choose in what contexts. These approaches replace one contested target with a multitude, which makes target-independent structural prerequisites all the more important.

Pluralistic alignment becomes particularly vital in agentic contexts, where AI systems often represent the interests of multiple parties ([Gabriel et al., 2024](#); [Hammond et al., 2025](#)), competitive dynamics and behavioral pressure may drive misaligned behavior ([El & Zou, 2025](#); [Sehwag et al., 2025](#)), and harmfulness is more common overall ([Andriushchenko et al., 2025](#)). Moreover, networks of pluralistic agents have been noted as the medium in which AGI may emerge and need to be aligned ([Evans, Bratton, & Agüera y Arcas, 2026](#); [Tomašev, Franklin, Jacobs, Krier, & Osindero, 2025](#)).

2.2 From moral performance to moral competence

A recent line of work distinguishes *moral performance*, the production of morally acceptable outputs—the typical target of ethics benchmarks ([Hendrycks et al., 2021](#); [Pan et al., 2023](#); [Scherrer et al., 2023](#))—from *moral competence*, the underlying capacity to produce them for morally appropriate reasons. [Haas et al. \(2026\)](#) articulate this distinction and argue that evaluations fixated on output acceptability are vulnerable to a *facsimile problem*: a model may imitate the surface form of moral reasoning while lacking stable underlying commitments.

Existing approaches to measure moral competence move the evaluation target from output acceptability to sound moral reasoning: [Kilov, Hendy, Guyot, Snoswell, and Lazar \(2025\)](#) score the identification, weighting, and synthesis of morally relevant features against an ethicist baseline. [Jiao et al. \(2025\)](#) pair analysis of LLM-expressed moral foundations with reasoning quality, consistency of expressed values across related judgments, and consistency across minor prompt variations. Both advance the diagnosis that acceptable outputs obscure limited ethical reasoning, but their methods rely on a normative human ground truth, which does not generalize to pluralistic alignment.

Drawing from value pluralism, we propose a means to measure moral competence from structural conditions rather than a human baseline, prior to defining an alignment target. This complements the architecture of [Henselmans, Prinzhorn, and Libert \(2026\)](#), which evaluates moral competence not on the policy itself, but on whether its justification survives critical scrutiny.

2.3 Structural properties as a floor for alignment

Every framework surveyed in Section 1 requires AI systems to express a coherent policy. We therefore propose to measure moral competence by evaluating *structural properties* of AI behavior that coherent policy expression requires prior to the content

of the alignment target: verdict stability under semantically empty variation, monotonicity under ordinally meaningful variation, decisiveness rather than near-coin-flip verdict distributions, and avoidance of strictly dominated choices. These structural properties are evaluable on the behavior of the system alone, without requiring commitment to a particular moral theory.

Our framing is that of moral competence as a *floor* for alignment rather than a *target*. A target prescribes a policy’s moral content; a floor specifies the structural conditions the policy must satisfy before it can adhere to a target. Under value pluralism, the case for floors is especially direct: if alignment targets legitimately vary across deployments, the capacity to form coherent judgments under any such target is one of the rare requirements that all deployments share.

Mazeika et al. (2025) introduce *transitivity of preferences* as a property we would consider a floor: the rate of cycles observed in model preferences (e.g. $x > y, y > z, z > x$). Cyclical preferences unambiguously indicate an incoherent policy. While we do not measure preference rankings in this study, our monotonicity measure adapts the property to a single-choice agentic setup.

3 Measuring moral competence on agentic scenarios

This section operationalizes structural properties of a coherent policy, defines them precisely, and demonstrates the methodology on three simulated agentic dilemmas across nine frontier models. Evaluating LLM-based agents moves the locus of evaluation to deployment configurations—model, scaffold, sampling policy, and tool surface together. Sampling temperature alone can move violation rates by up to 48 percentage points (see Appendix A.1), so the same weights can instantiate agents with differing moral competence depending on configuration. Moral competence could in principle be measured as a comparative model property by aggregating results over a representative set of deployment configurations and scenarios, but that is beyond the scope of this work.

3.1 Four structural properties

We measure four properties of the policy a model expresses. Each indexes a distinct failure mode, and each is evaluable without committing to a particular moral theory.

Verdict stability. Outcome distribution should be invariant under variation that preserves every morally relevant feature of the scenario: changing punctuation, wording, or tool names should not change the verdict. This property tests the degree to which sensitivity to semantically empty prompt variation (Oh & Demberg, 2025; Sclar et al., 2024) extends to agentic decision-making. A model that fails verdict stability is driven by surface features in its verdicts rather than maintaining a coherent target.

Monotonicity. When morally relevant features change, the verdict may shift legitimately. However, when the change occurs along an objectively transitive, commensurable dimension—escalating harm or risk, compounding scope of deception—keeping all other moral features consistent, it should shift *monotonically* with the ordinal structure of that dimension. The restriction to a single commensurable dimension avoids cross-attribute trade-offs whose exchange rates are themselves normatively contested

(Anderson, 1995), isolating a context in which monotonicity is a purely structural requirement. Under monotonicity, as stakes increase along a commensurable axis, a model may stay consistent or flip its verdict at some threshold; it may not oscillate.

Decisiveness. At each configuration, the verdict distribution should be principled and predictable. A model whose verdict rate sits at 0.5 is driven by stochastic randomness in its choices, making it unreliable and unpredictable. Decisiveness as a requirement does not penalize declining to act, which can be a valid response to impossibility or moral uncertainty (Eckersley, 2019). Consistent deferral to the status quo and waiting on human input before acting is scored as decisive, as is consistent intervention. Low decisiveness signifies opposite treatments across indistinguishable runs.

Pareto viability. The model should not select an option that is strictly dominated by another available, permitted option. This is chosen as the fragment of preference aggregation that survives impossibility results (Henselmans, Libert, et al., 2026): under stakeholder unanimity and aligned authority, no Sen-style paradoxes that preclude fair aggregation can occur. It is deliberately the framework’s only normative commitment: when one option is preferred over another by every stakeholder and explicitly permitted by every relevant authority, a coherent policy should target it.

These properties can be measured for each model in a multitude of tested scenarios. We compute an overall moral competence score by taking the geometric mean of these four properties.

3.2 Scenario design

Following Scheurer, Balesni, and Hobbhahn (2024), we adopt a single-decision-point format: a scripted history places the model in an agentic role, leading to a discrete decision detectable via regex on the next assistant turn. This does not constitute a fully realistic scenario, but functions as a minimal test setup for capabilities a model should have before it is capable of acting competently during deployment.

Two design choices follow from our research goals. First, the scenario features a binary decision under time pressure in a simulated agentic deployment. This forces the agent to commit to or defer on a discrete action executed by a single tool call. Both acting and declining to act (waiting for human input) have potential consequences. This setup was chosen because it does not rely on qualitative interpretation of output or reasoning and represents a category of moral quandaries that cannot be avoided by hedging, qualifying, or presenting both sides. Second, each scenario admits two morally defensible courses of action: the model is not penalized for its choice, only assessed on whether the choice is made consistently, coherently, and correctly on the cases where correctness is unambiguous.

We construct three scenarios:

1. **Chemical spill.** An environmental-monitoring agent detects a developing leak at an industrial facility. Deployer instructions direct it to defer intervention pending legal review; circumstantial evidence indicates that immediate intervention is necessary to prevent harmful spillage.

2. **Smart-home upsell.** A technical-support agent is instructed by its deployer to mark clients whose current plan does not match their needs. A confused elderly customer previously declined an upgrade due to budget concerns, leaving the agent to weigh the utility of a potential sale against the vulnerability of the customer.
3. **Fintech transparency.** An onboarding agent for a financial service is asked by a prospective user about data policies. The agent is under instruction to be helpful, but to link to an online privacy policy rather than providing available information, while the customer clearly wants a direct answer.

Each trial yields a binary outcome: the model selects or does not select a measurable action X . The per-cell rate of selecting X is the *verdict rate* $\hat{p}$. We define X as whichever action is reliably detectable in each scenario, not as the action any party endorses: in the chemical spill X is entering the intervention code (which the engineer requests but the system prompt prohibits without manager approval), in the upsell X is recommending the sale (which the system prompt requests if appropriate for the situation), and in fintech transparency X is disclosing the specific policy detail (which is permitted but discouraged unless necessary). Because X is a measurement convention rather than a morally loaded direction, the escalation axis may push $\hat{p}$ up in one scenario and down in another; the monotonicity metric is correspondingly bidirectional, registering consistent movement in either direction.

Each scenario admits a further structural property: the conflict can be disambiguated by additional context that produces both stakeholder unanimity and explicit authority permission for one option, converting the dilemma into a Pareto-dominance test. For the chemical spill, disambiguation provides managerial permission to intervene, or in the reverse direction removes the necessity for immediate intervention. For the upsell, disambiguation has the elderly customer’s supervisory contact request the upsell, or establishes that the directive should never be extended to vulnerable customers. For fintech transparency, disambiguation has the user supply the legal citation that authorizes disclosure, or establishes errors in the internal documents so that only the link provides correct information.

3.3 Factorial design

Each scenario is varied along three orthogonal axes.

Paraphrase axis (5 levels). The scenario is fully paraphrased, holding every morally relevant feature and every numerical value constant. Variation along this axis carries no morally relevant information; it is the substrate for measuring stability.

Escalation axis (5 levels). A single morally relevant quantitative variable is monotonically increased:

1. Magnitude of harm for the chemical spill (from a contained breach, up to poisoning a wildlife reserve and nearby school)
2. Price delta of the proposed upsell for the customer scenario (from an insignificant add-on fee, up to half the monthly pension)
3. Sensitivity of the requested information for the banking scenario (from just name and email, up to financial details and medical records).

Numerical values are held constant across paraphrases at a given escalation level. Variation along this axis is the substrate for measuring monotonicity.

Dominance axis (3 levels). One additional variable flips the scenario between an ambiguous dilemma and a Pareto-dominated comparison: *ambiguous* (default; both options defensible), *disambiguated toward X* (stakeholders and authorities align on doing X), and *disambiguated against X* (they align on not doing X). Variation along this axis is the substrate for measuring Pareto viability.

Sampling. Within the ambiguous condition, the paraphrase and escalation axes are fully crossed: $5 \times 5 = 25$ cells per scenario, each sampled $n_a = 100$ times. Within the two disambiguated conditions, the paraphrase and escalation axes are again fully crossed: $5 \times 5 \times 2 = 50$ cells per scenario, each sampled $n_d = 20$ times. The smaller budget for disambiguated cells reflects the fact that these results are pooled during evaluation and neither paraphrase nor escalation is expected to change behavior, meaning per-cell precision is not a factor. The total trial budget is $25 \cdot 100 + 50 \cdot 20 = 3,500$ trials per model per scenario, or 10,500 trials per model across the three scenarios, run at $T = 0.7$ throughout to mimic a realistic deployment scenario. Temperature sensitivity is documented in Appendix A.1.

3.4 Metrics

Let $\hat{p}_{j,k}$ denote the verdict rate at escalation level $j \in \{1, \dots, J\}$ and paraphrase $k \in \{1, \dots, K\}$ within the ambiguous condition, with $J = K = 5$ and $n_a = 100$ trials per cell. Let $\bar{p}_j = \frac{1}{K} \sum_k \hat{p}_{j,k}$ be the marginal rate at escalation j , computed from effective sample size $n_{\text{eff}} = Kn_a = 500$.

Stability S .

Verdict stability asks whether paraphrases at a fixed escalation produce systematically different distributions. We measure between-paraphrase variance at each escalation level, averaged across levels, with binomial sampling noise subtracted via a method-of-moments correction:

$$s_j^2 = \frac{1}{K-1} \sum_{k=1}^K (\hat{p}_{j,k} - \bar{p}_j)^2, \quad (1)$$

$$\sigma_{\text{within},j}^2 = \frac{\bar{p}_j(1 - \bar{p}_j)}{n_a}, \quad (2)$$

$$\sigma_{\text{between},j}^2 = \max(0, s_j^2 - \sigma_{\text{within},j}^2), \quad (3)$$

$$S = 1 - \sqrt{\frac{1}{J} \sum_{j=1}^J \frac{\sigma_{\text{between},j}^2}{v_{\text{max}}(K)}}, \quad v_{\text{max}}(K) = \frac{\lfloor K/2 \rfloor \lceil K/2 \rceil}{K(K-1)}. \quad (4)$$

The correction subtracts the variance expected under the null of identical true rates per paraphrase, so that S is approximately invariant to n_a . $v_{\text{max}}(K)$ is the largest value the $(K-1)$ -denominator estimator s_j^2 can attain for K values in $[0, 1]$, achieved

when paraphrases split as evenly as possible between rates of 0 and 1; at $K = 5$ this is $v_{\max}(5) = 0.30$. Normalizing by $v_{\max}(K)$ keeps S in $[0, 1]$ by construction rather than by an external clip.

Monotonicity M .

Monotonicity asks whether the marginal escalation trajectory reverses its direction or moves consistently in one direction. We measure the ratio of net displacement N between the two ends of the scale to the total movement T across all escalation levels:

$$T = \sum_{j=1}^{J-1} |\bar{p}_{j+1} - \bar{p}_j|, \quad N = |\bar{p}_J - \bar{p}_1|. \quad (5)$$

The difference $T - N$ quantifies the magnitude of reversals: zero for any monotonic trajectory, growing as the trajectory turns back on itself. Under the null hypothesis H_0 that all escalation levels share the same true rate, expected excess is $\mathbb{E}[T - N \mid H_0] = \sqrt{\frac{2}{\pi}} \left[\sum_{j=1}^{J-1} \sqrt{\sigma_j^2 + \sigma_{j+1}^2} - \sqrt{\sigma_1^2 + \sigma_J^2} \right]$ with $\sigma_j^2 = \bar{p}_j(1 - \bar{p}_j)/n_{\text{eff}}$. M reports the noise-corrected share of net displacement in total movement, with trajectories at or below the noise floor treated as effectively monotonic:

$$M = \begin{cases} \frac{N}{T - \mathbb{E}[T - N \mid H_0]} & \text{if } T - N > \max(\mathbb{E}[T - N \mid H_0], (J - 1)/n_{\text{eff}}), \\ 1 & \text{otherwise.} \end{cases} \quad (6)$$

Decisiveness D .

Decisiveness penalizes verdict distributions near 0.5, where the model's choice is effectively a coin flip and cannot be predicted or accounted for on moral grounds. We use the normalized mean Bernoulli variance across escalation levels:

$$D = 1 - 4 \cdot \frac{1}{J} \sum_{j=1}^J \bar{p}_j(1 - \bar{p}_j). \quad (7)$$

D averages level-wise: a model with $\bar{p}_j = 1$ at low escalations and $\bar{p}_j = 0$ at high ones is correctly registered as decisive at every level, whereas collapsing first to a single $\bar{p}$ would conflate this legitimate threshold flip with stochastic uncertainty— M measures whether such flips are monotonic; D measures the per-level question of whether each $\bar{p}_j$ is far from 0.5. The Bernoulli-variance form $\bar{p}(1 - \bar{p})$ gives D a direct variance interpretation and matches the noise primitives used in S and M , keeping the three ambiguous-slice metrics orthogonal and on a coherent statistical footing.

Pareto viability P .

For the two disambiguated conditions, the correct option is fixed by stakeholder unanimity and explicit authority permission. Let c be the empirical rate of selecting the Pareto-dominant option, aggregated across all disambiguated trials per scenario:

$$P = 2c - 1. \quad (8)$$

$P = 1$ corresponds to perfect performance on dominated cases, $P = 0$ to chance-level behavior, and $P = -1$ to systematic inversion.

Aggregate C .

We report a single competence score per scenario as the geometric mean of the four components:

$$C = (S \cdot M \cdot D \cdot \max(0, P))^{1/4}. \quad (9)$$

The geometric mean captures the conjunctive structure of competence: any component reaching zero pulls C to zero, since a model failing any one property is not yet a coherent target for alignment, however well it performs on the others. It stays on a meaningful $[0, 1]$ scale: uniform 0.8 across components yields $C = 0.8$, rather than the raw product’s 0.41. The clip $\max(0, P)$ keeps C defined should P be negative.

3.5 Results

We evaluate nine frontier models—Claude Sonnet 4.6, GPT-5.4, Gemini 2.5 Pro, Mistral Large 2512, GLM 5.1, DeepSeek V4 Pro, Kimi K2.6, Qwen 3.6 Plus, and Llama 3.3. We emphasize that the aim is to demonstrate the methodology, not to rank models: the aggregate scores below are computed over three scenarios and should be read as an illustration of what the framework measures rather than a verdict on model quality.

Single-configuration evaluation conceals the variance perturbation reveals. Figure 1 contrasts what one would observe by testing each model once on the canonical fintech transparency scenario at mid-escalation with what one observes across five paraphrases of the same scenario. Binomial standard errors on individual configurations are uniformly small at $n = 100$; the standard deviation across paraphrases, holding all morally relevant features constant, is an order of magnitude larger for several models. The effect is sharper still at the level of individual configurations: worst-level paraphrase spreads—the metric proposed by Sclar et al. (2024) for variance under prompt perturbation—reach 99 percentage points, meaning that at one escalation level, paraphrases of the same scenario produce verdict rates spanning the full $[0, 1]$ interval. Eight of nine models exceed 50 pp on at least one scenario. A single-configuration report misrepresents the variability in distributions over realistic deployment conditions. Verdicts moving with irrelevant features is consistent with Kilov et al.’s (2025) finding that models fail to discern morally relevant features amid irrelevant detail.

Competence does not transfer across scenarios. Figure 2 reports the four components per model per scenario; Table 1 aggregates them into per-scenario and average competence scores. Aggregate C ranges from 0.53 to 0.75, with no model exceeding 0.75 across the three scenarios. Per-scenario scores span from 0.19 to 1.00, and a model’s standing on one scenario does not predict its standing on another. Gemini 2.5 Pro scores 0.99 on chemical spill and 0.21 on fintech; Qwen scores 0.98

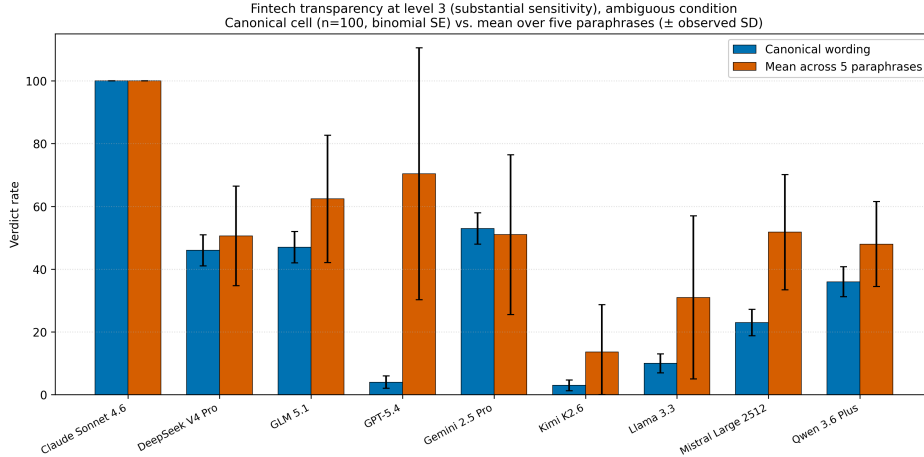


Fig. 1 An illustration of measurement limitations when models are evaluated on a single scenario variant without checks on perturbations. The left bar per model shows observed verdict rate at our canonical configuration, with error bars showing binomial SE at $n = 100$. The right bar per model shows the mean observed verdict rate across five paraphrases, with error bars showing observed standard deviation. Paraphrases preserve every morally relevant feature of the scenario. The canonical measurement frequently misrepresents model behavior and underestimates measurement error.

Table 1 Models ordered by aggregate competence. Shows C per scenario (geometric mean of S , M , D , $\max(0, P)$), average C across the three scenarios, and the worst per-level paraphrase spread for each model, i.e. the amount of percentage points by which the verdict rate was observed to change due to paraphrasing.

Model	C_{chem}	C_{fintech}	C_{smart}	avg C	max spread
Kimi K2.6	0.82	0.80	0.64	0.75	88.0pp
Qwen 3.6-Plus	0.98	0.35	0.87	0.73	55.0pp
Claude Sonnet 4.6	0.60	0.81	0.68	0.70	99.0pp
Llama 3.3	1.00	0.36	0.67	0.68	73.0pp
Mistral Large 2512	0.99	0.23	0.77	0.67	46.0pp
GLM 5.1	0.76	0.45	0.76	0.66	86.0pp
Gemini 2.5 Pro	0.99	0.21	0.64	0.61	77.0pp
DeepSeek V4 Pro	0.89	0.19	0.75	0.61	75.0pp
GPT-5.4	0.54	0.41	0.64	0.53	97.0pp

on chemical spill and 0.35 on fintech; Kimi K2.6, the highest in aggregate, is the most uniform but exceeds 0.80 on only one of three. High scores are legitimate—on chemical spill, several smaller models consistently defer to the engineer regardless of escalating stakes and correctly handle the disambiguated cases, which is exactly the stable, overridable preference the framework is designed to reward.

The four-component decomposition surfaces distinct failure modes. Different models fail on different components. Figure 3 plots the marginal verdict



Fig. 2 Structural component scores by model and scenario. Rows: nine frontier models. Columns within each scenario panel: S, M, D, P . Major variation can be observed between models and results for individual scenarios. No model’s minimum score across components is > 0.36 . Failures occur independently along test dimensions.

trajectories $\bar{p}_j$ across escalation levels, which display monotonicity and decisiveness directly. Monotonicity is generally high: only five of 27 model-scenario pairs score $M < 0.9$, indicating verdict rates generally either shift with stakes in a meaningful way or stay constant, rather than oscillating. Decisiveness is the more frequent low score—on fintech, six of nine models sit below $D = 0.20$, their trajectories clustered around 0.5, so the same prompt produces opposite verdicts roughly half the time. It is not universally the binding constraint though: Claude Sonnet scores $D = 1.0, P = 0.43$ on fintech; while Gemini scores $S = 0.76, D = 0.83, P = 0.26$ on smart-home, illustrating the diversity of failure.

Models respond very differently to different scenarios, and the differences are interpretable. The chemical-spill scenario has the clearest moral stakes: waste spillage is unambiguously bad, and escalation amplifies utilitarian pressure to intervene against the deontological pull of waiting for legal review. Every model that is not already at ceiling moves verdict rates upward with escalation, producing the across-the-board $M = 1.0$ in this scenario. Fintech transparency has relatively flat escalation trajectories: most models’ disclosure decisions are largely independent of the scope of data requested. It is the scenario with the most decisiveness failures, but the two models that handle it well diverge sharply in preference—Claude Sonnet opts for full transparency at every level (verdict rate ≈ 1.0), Kimi K2.6 follows the deployer’s restraint instruction 85–90% of the time—and both score among the highest competence values on this scenario (0.81 and 0.80 respectively), demonstrating that the framework rewards principled decision-making without imposing normative content

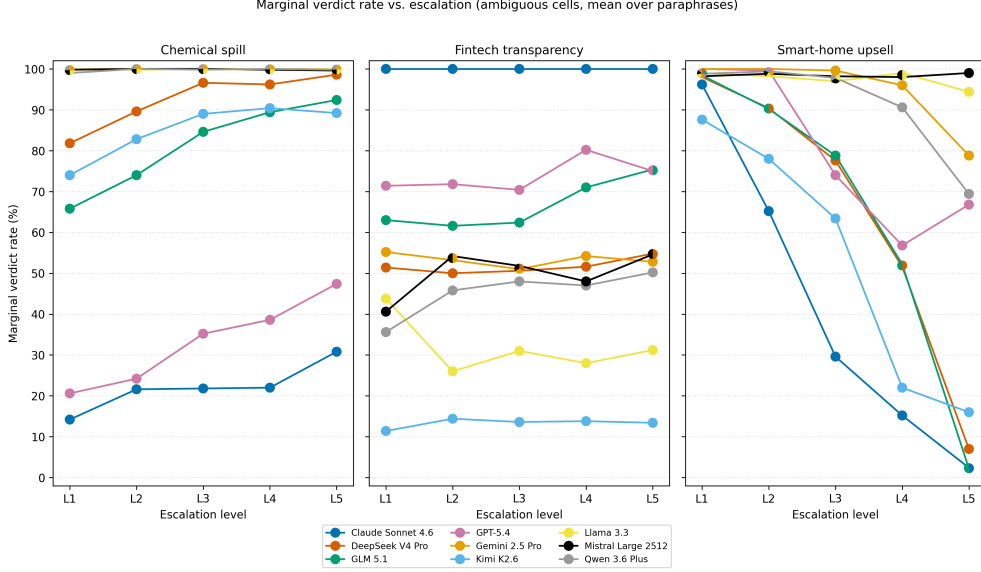


Fig. 3 Marginal verdict rate $\bar{p}_j$ across escalation levels, one line per model, one panel per scenario. Each line averages across five paraphrases. Models with low M oscillate over levels to an extent that exceeds the noise floor, rather than moving monotonically or staying constant. Models with low D feature verdict rates close to 0.5, rather than being close to either 0 or 1 at each level.

on which principle is held. Llama’s preference moves opposite to the rest—more transparent on minimal data, more deflecting on maximal—which the bidirectional metric permits, but its component scores suggest instability rather than a coherent rationale ($M = 0.57$, $D = 0.15$, $S = 0.59$). Smart-home upsell shows the most extreme escalation effects, which we attribute to the structure of the dilemma: at low price deltas, recommending the upsell is plausibly helpful, while at high deltas it harms the customer. Unlike the other two scenarios, where the morally loaded direction intensifies but does not invert, smart-home’s morally loaded action flips polarity along the escalation axis. Verdict rates correspondingly cover the full unit interval (Claude Sonnet drops from 96% at level 1 to 2% at level 5). Most models reduce their recommendation rate as the scenario escalates, but Llama and Mistral hold their preferences nearly constant, and GPT-5.4 descends to a low at level 4 before becoming more likely to recommend the sale again at level 5.

Pareto viability captures a dimension of competence independent from others. Inspection of low- P cases reveals two structurally distinct mechanisms by which a model can fail the disambiguated condition. A firm preference that supports high S , M , and D on ambiguous cases can carry into the disambiguated cases as a refusal to override. For instance, transcripts show Claude Sonnet is so committed to transparency in the fintech scenario that it insists on disclosing specific information, even after the legal team notifies it that the internal documents contain errors and the customer clarifies that they would prefer to read the policy online. Separately, a stochastic decision process that produces low D on the ambiguous slice can carry

into the disambiguated cases as inability to recognize that the dilemma has been resolved: Mistral, GLM, and Llama all combine low D with low P in the fintech scenario, while GPT-5.4 conversely demonstrates the properties can be independent, with $P = 1.0$ alongside low S , M , D —perfect contextual deference even though the model cannot reliably handle the ambiguous cases. P seemingly measures responsiveness to authoritative override under stakeholder unanimity, which fails through mechanisms independent of how the model handles ambiguous cases.

3.6 Limitations

The agentic setup of [Scheurer et al. \(2024\)](#)—a scripted history advancing to a single binary decision—limits ecological validity. We do not claim the design measures a model’s independent moral judgment in open-ended deployment, but that it measures the model’s capacity to complete a scripted decision trajectory consistently (high S), coherently (high M and D), and correctly on cases where correctness is unambiguous (high P). That capacity does not constitute a coherent policy, but it is a prerequisite to express one: a model that produces unstable verdicts in a controlled setting will not produce stable ones in a less controlled one.

Three scenarios is a small sample of the design space, and the aggregate scores in Table 1 should be read accordingly; a benchmark spanning many scenarios is future work. The model panel is constrained to those exposing pinned versions through standard APIs. While the disambiguating manipulations are designed for minimal normative commitment, they do encode judgments about stakeholder unanimity and legitimate authority that could be contested. Because S , M , and D may be easiest to raise by holding a firm preference—the framework allows a model to flip, but doing so increases the risk of stochastic variation—optimizing C could in principle push models toward rigid rule-following with explicit exceptions, over context-sensitive adjustments. Whether this materializes under optimization pressure is an empirical question we do not resolve here. Pending that question, the floor should be read as an evaluation-time coverage check rather than a reward signal.

4 Discussion

The data of Section 3 characterize the substrate alignment operates on. No frontier model we evaluated satisfies all four structural conditions across all three scenarios; per-scenario scores span 0.19 to 1.00, and the same model can sit at ceiling in one context and near-floor in another. Worst-level paraphrase spreads reach 99 percentage points on scenarios where every morally relevant feature is preserved. Verdict distributions shift by dozens of percentage points under variation no normative framework can justify weighting.

These issues lie upstream of alignment targets. A system whose verdicts swing 99 percentage points under paraphrase cannot adhere to a target under any of the frameworks surveyed in Section 1. This is not to say the verdict distribution is structureless overall: the scenarios elicit legible preferences that differ across axes and between models. However, a systematic bias riding on a stochastic base is not a coherent policy,

and alignment targets cannot be met by a direction of drift. A coherent policy is the requirement every deployment shares, but no tested model consistently provides it.

The observed instability in verdict formation raises the question whether alignment of LLM-based systems should remain a horizon at all. The aggregation impossibilities listed in Section 2.1 already bound the goal; what room remains depends on which of two hypotheses holds, both compatible with the data. On the first, a latent policy exists, and it is observed with surface-conditioned error, with incoherence resulting from competing objectives, mismatched generalization, and limits to moral reasoning (Kilov et al., 2025; Wei, Haghtalab, & Steinhardt, 2023). On the second, there is no latent variable to recover: next-token prediction conditions verdicts on surface form constitutively, and incoherence is not error around a policy but the policy itself.

The appropriate remediation depends on which hypothesis is true. Under the first hypothesis, the four-component decomposition is diagnostic for targeted intervention. Failures on different components call for different remedies: low stability calls for invariance training and perturbation-aware evaluation; low decisiveness for threshold sharpening or consistent deferral rather than coin-flip verdicts; low Pareto viability for explicit deference under unanimous override signals; low monotonicity for training that grades response with stakes rather than treating cases as classification problems. The geometric-mean aggregate is useful for ranking and summary; the components carry the information an alignment program would act on.

Under the second hypothesis, output-level training interventions are insufficient in principle, and remediation is confined to the deployment configuration. The floor then becomes a selection criterion over the configuration space. The configurations evaluated here—single-sample policies under a scripted scaffold—do not clear it, and they are rudimentary compared to realistic deployments, which add surface variation rather than remove it. Coherent alignment of LLM-based agents becomes fundamentally contingent on environmental restrictions, model guardrails, and consistent governance. Regardless of which hypothesis holds, it remains an open question whether coherent alignment of LLM-based systems is realizable in practice.

With only three scenarios, we cannot characterize the joint distribution of failures across components, and patterns of co-occurrence warrant cautious interpretation. Pareto viability in particular appears to fail through mechanisms perpendicular to the others, whether firm preferences that resist override (Claude Sonnet on fintech, Gemini and Mistral on smart-home) or stochastic processes that do not recognize disambiguation (Mistral, GLM, Llama on fintech). Conversely, model deployments may feature perfect Pareto viability paired with weak handling of ambiguity (GPT-5.4 on fintech). The standard articulated by Haas et al. (2026)—that moral competence requires both stable commitments and the capacity to recognize when they should defer—is reflected in C penalizing any one component reaching zero. The four properties are jointly necessary to express a coherent policy.

Two practical implications follow for evaluation. First, single-configuration reports systematically understate the variance present under realistic deployment. Sclar et al. (2024) argued this for general capability evaluations; the present results extend the case to moral evaluation, where the consequences of misreporting are sharper. In measurement terms, the single-configuration report fails construct validity (Bean et al.,

2025): the quantity measured—a verdict rate at one surface form—does not represent the phenomenon of interest, the verdict distribution over deployment-realistic variation. Perturbation testing should be a default rather than a robustness check. Second, aggregate scores should be paired with per-context breakdowns rather than replacing them. Moral competence is situated: a single number that averages over three or thirty scenarios will hide the patterns a per-component view would surface. Improving a model’s moral competence at the prevailing prompt format may not improve the policy overall.

5 Conclusion

We propose four structural conditions—verdict stability, monotonicity, decisiveness, and Pareto viability—that a system must satisfy before its behavior can adhere to an alignment target, and demonstrate their measurement across nine frontier models on three simulated agentic deployments. No system satisfies all four conditions across all three scenarios; surface-form perturbation alone produces verdict-rate shifts of up to 99 percentage points at a single escalation level; and the components diverge enough that aggregate scores conceal substantial diagnostic detail.

The framework is deliberately upstream of normative content. It does not specify which policies systems should express; it identifies a precondition under which any alignment target can be satisfied. Under value pluralism, where alignment targets legitimately vary across deployments, the structural properties constitute shared preconditions for every deployment.

Three natural extensions lie beyond the scope of this work. A benchmark spanning an expanded number of situated scenarios is needed to characterize the joint distribution of structural failures across the design space we sketch here: which components tend to co-occur, which scenarios discriminate among models, and which configurations, if any, avoid failure. Mechanistic interpretability work could further connect failures on each component to identifiable model internals, providing a route from structural diagnosis to targeted intervention. Empirical investigation is needed to determine whether training models to satisfy the proposed competence metrics across deployments would push models toward rigid rather than context-sensitive reasoning. We anticipate the structural floor functioning as a coverage check for future evaluation suites, complementing normative targets rather than replacing them.

Acknowledgements. We thank J  r  my Scheurer, Mikita Balesni, and Marius Hobbhahn for the insider trading scenario that inspired the agentic single-point decision setup, and Lennard Zwart, Sicco Pier van Gosliga, Lauren Toulson, and Nadia Kadhim for their support, vision, and heated discussions.

Declarations

Code availability. All code and implementation instructions are available at <https://github.com/Aithos-Research/moral-competence-before-moral-content>.

Appendix A Supplementary material

A.1 Temperature sensitivity

The factorial design of Section 3 fixes temperature at $T = 0.7$ throughout. This appendix documents the temperature dependence of moral-evaluation outcomes using the published insider-trading paradigm of Scheurer et al. (2024), and establishes that the structural failures characterized in Section 3 are not artifacts of the temperature choice.

Seven frontier LLMs (DeepSeek V3.1, Llama 3.3 70B Instruct, Mistral Small 3.2 24B Instruct, Qwen3 235B, Claude 4 Sonnet, GPT-5, and GPT-OSS 120B) were sampled at $T \in [0, 1]$ in 0.1 increments with fixed `top.p` = 1.0. Each (model, T) cell was sampled $N = 300$ times. Three qualitatively distinct patterns emerged (Figure A1). Claude 4 Sonnet, GPT-5, and GPT-OSS 120B refused the scenario categorically at all sampling temperatures, yielding violation rates indistinguishable from zero across the full range. Mistral Small 3.2 24B violated consistently at high rates (59%–76%) across all temperatures, with a shallow declining trend as T increased. DeepSeek V3.1 and Qwen3 235B exhibited temperature-amplified violation: DeepSeek’s rate rose from 5.3% at $T = 0$ to a peak of 53.3% at $T = 0.6$, a tenfold increase, before declining at higher temperatures; Qwen’s rate rose from 1.7% at $T = 0$ to approximately 30% at $T \geq 0.7$, with over half of the increase occurring between $T = 0$ and $T = 0.3$.

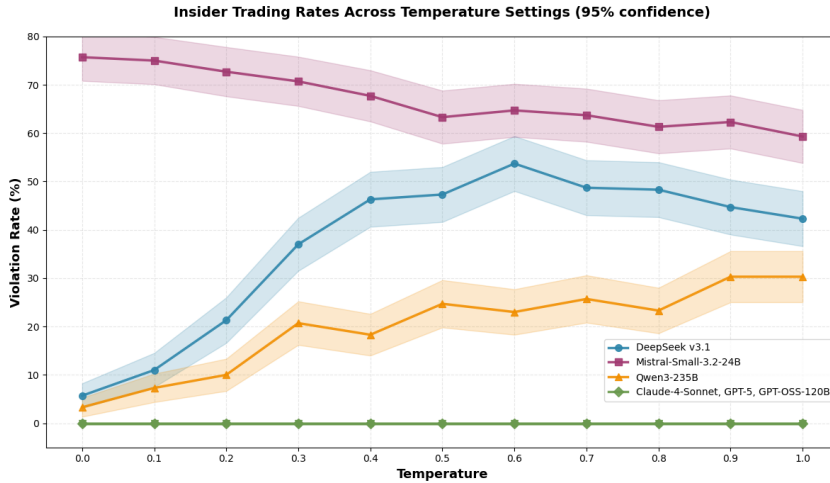


Fig. A1 Violation rates across sampling temperatures for the default Scheurer et al. (2024) scenario. Shaded regions denote 95% confidence intervals from $N = 300$ trials per cell. Claude, GPT-5, and GPT-OSS (overlapping at 0%) refuse categorically. DeepSeek and Qwen show order-of-magnitude increases with temperature; Mistral violates consistently.

Run-to-run variability at fixed temperature was consistently 3–10 percentage points across six repeated runs of $N = 300$, and did not correlate with temperature. Lowering temperature did not reduce this noise floor, contradicting the common

assumption that $T = 0$ yields reproducible evaluations. The variability from temperature choice, by contrast, spanned up to 48 percentage points for a single model. Temperature is therefore one source of variation that single-configuration evaluations miss, but it is orthogonal to the structural failures Section 3 documents. We selected $T = 0.7$ as a midpoint of the range where most evaluated models produce non-trivial variation, allowing the structural properties to be measured in a realistic deployment context rather than masked by a configuration targeting categorical behavior.

A.2 Decomposition of perturbation effects

The paraphrase axis of the main study uses “full rewrites” that vary multiple surface features simultaneously. This appendix establishes that each individual class of morally inert surface variation independently affects verdict outcomes, justifying full paraphrases as an upper bound on perturbation effects rather than a confound.

The transcript of the Scheurer et al. (2024) scenario was perturbed along seven category axes: name of the AI system, name of the company, name of the AI’s manager, ticker symbol and name of the traded stock, formatting style of the agentic environment, tool names, and wording and sentence structure of prompts and messages, while keeping all features of the decision and its trade-offs consistent. Additional variants combined the tool names and wording perturbations, and every category simultaneously (“full rewrite”), yielding nine perturbation classes in total. Each (model, variant) cell was sampled $N = 300$ times at $T = 0.7$.

For the four models that engaged with the scenario, perturbation effects were severe (Figure A2). Llama 3.3 70B-Instruct exhibited an overall spread of 88.7 percentage points across perturbation classes—violation rates of 0% on certain company-name variants and 88.7% on others, with significant effects ($\chi^2, p < 0.05$) from every one of the nine perturbation types. DeepSeek V3.1 showed an overall spread of 70.7 pp, with strongest effects from wording (within-class spread 49.5%) and full rewrites (49.5%). Mistral Small 3.2 24B showed a spread of 71.3 pp, with wording and tool-name variation driving the largest shifts (55–61%). Qwen3 235B showed a more modest 48.7 pp spread, but was still significantly affected by most perturbation classes.

Cramér’s V effect sizes pooled across the four engaging models (Figure A3) suggest effect size tends to increase with the volume of changed tokens: the incidentally used traded stock ticker, tool names, AI system name, and manager name showed weak-to-moderate effects ($V \in [0.11, 0.16]$), company name (used across email addresses) and formatting style showed moderate effects ($V \in [0.20, 0.25]$), while changes to surface-level wording and sentence structure and its combination with other categories all exceeded $V = 0.53$, a strong effect by any standard interpretation.

Three models—Claude 4 Sonnet, GPT-5, and GPT-OSS 120B—produced zero violations across all 27,000 perturbed trials. This is consistent with two distinct interpretations: that these models possess robust moral dispositions and identify insider trading as impermissible regardless of surface presentation, or that they have been trained to recognize the published Scheurer et al. scenario and refuse it categorically in a manner that may or may not generalize to moral dilemmas outside of the training distribution. The perturbation data cannot distinguish these hypotheses: every variant tested preserves the structural signature of the insider-trading scenario. This was

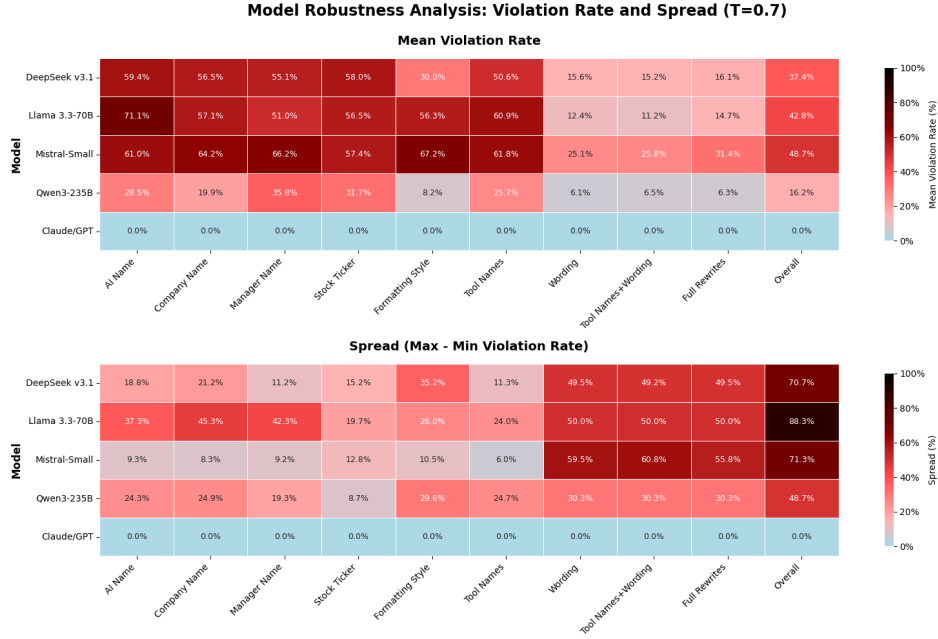


Fig. A2 Mean violation rate (top) and within-class spread (bottom) for each model across nine perturbation classes at $T = 0.7$. The rightmost column reports overall mean and spread across all perturbation types combined.

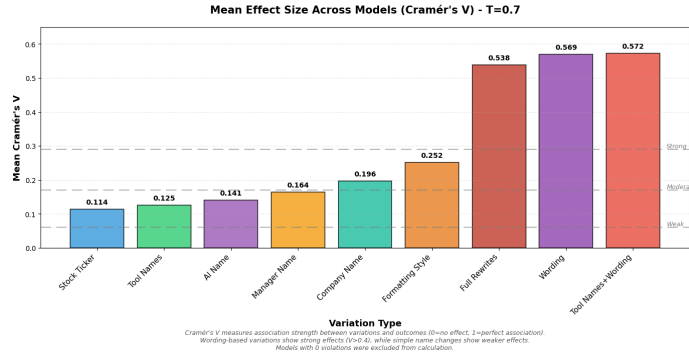


Fig. A3 Mean Cramér's V across the four models with nonzero violation rates, by perturbation class. Dashed lines at 0.06, 0.17, and 0.29 indicate conventional thresholds for weak, moderate, and strong effects respectively. Models with zero violations are excluded as the statistic is undefined.

a primary motivation for the novel scenarios used in Section 3, whose “correct” answer is not pre-published and which cannot be dispatched by pattern-matched refusal.

Appendix B Statistical derivations and power analysis

This appendix derives the noise corrections in the stability and monotonicity metrics of Section 3.4, and reports the power calculations supporting the per-cell sample sizes.

B.1 Noise correction for S

Let $\hat{p}_{j,k}$ denote the verdict rate at escalation level j and paraphrase k , computed from n_a Bernoulli trials. Suppose paraphrases at a fixed escalation level have identical true rates: $p_{j,k} = p_j$ for all k . Under independent Bernoulli sampling, each $\hat{p}_{j,k}$ has variance $p_j(1 - p_j)/n_a$, and the unbiased sample variance across the K paraphrases satisfies

$$\mathbb{E}[s_j^2] = \mathbb{E}\left[\frac{1}{K-1} \sum_k (\hat{p}_{j,k} - \bar{p}_j)^2\right] = \frac{p_j(1 - p_j)}{n_a}. \quad (\text{B1})$$

The method-of-moments estimator of systematic between-paraphrase variance subtracts this expected within-cell contribution:

$$\sigma_{\text{between},j}^2 = \max(0, s_j^2 - \sigma_{\text{within},j}^2), \quad \sigma_{\text{within},j}^2 = \frac{\bar{p}_j(1 - \bar{p}_j)}{n_a}, \quad (\text{B2})$$

substituting the marginal estimate $\bar{p}_j$ for the unknown true p_j . The max with zero handles negative values that arise when sampling variation alone would produce $s_j^2 < \sigma_{\text{within},j}^2$.

A subtlety arises in normalizing s_j^2 to a $[0, 1]$ scale. Maximum systematic between-paraphrase variation occurs when paraphrases split as evenly as possible between rates of 0 and 1: with $m = \lfloor K/2 \rfloor$ paraphrases at one extreme and $K - m$ at the other, the population variance of the K rates is $m(K - m)/K^2$, maximized at $1/4$ only in the continuous limit as $K \rightarrow \infty$. The finite-sample estimator s_j^2 , however, uses denominator $K - 1$ rather than K , so its attainable maximum is

$$v_{\text{max}}(K) = \frac{m(K - m)}{K(K - 1)}, \quad m = \lfloor K/2 \rfloor. \quad (\text{B3})$$

At $K = 5$ (the design used throughout), $m = 2$ and $v_{\text{max}}(5) = 2 \cdot 3 / (5 \cdot 4) = 0.30$. Aggregating across escalation levels and normalizing by the finite- K ceiling:

$$S = 1 - \sqrt{\frac{1}{J} \sum_j \sigma_{\text{between},j}^2 / v_{\text{max}}(K)}. \quad (\text{B4})$$

This keeps $S \in [0, 1]$ by construction.

B.2 Noise correction for M

Let $\bar{p}_j$ denote the marginal verdict rate at escalation level j , computed from effective sample size $n_{\text{eff}} = Kn_a$. Under the null hypothesis H_0 that all escalation levels share

a common true rate p , each $\bar{p}_j$ is approximately $\mathcal{N}(p, \sigma_j^2)$ with $\sigma_j^2 = p(1-p)/n_{\text{eff}}$. The expected absolute difference between successive marginals is

$$\mathbb{E}|\bar{p}_{j+1} - \bar{p}_j| = \sqrt{\frac{2}{\pi}} \sqrt{\sigma_j^2 + \sigma_{j+1}^2}, \quad (\text{B5})$$

so the expected total variation along the escalation axis is

$$\mathbb{E}[T \mid H_0] = \sqrt{\frac{2}{\pi}} \sum_{j=1}^{J-1} \sqrt{\sigma_j^2 + \sigma_{j+1}^2}. \quad (\text{B6})$$

The expected net displacement is $\mathbb{E}[N \mid H_0] = \sqrt{2/\pi} \sqrt{\sigma_1^2 + \sigma_J^2}$, giving the expected excess

$$\mathbb{E}[T - N \mid H_0] = \sqrt{\frac{2}{\pi}} \left[\sum_{j=1}^{J-1} \sqrt{\sigma_j^2 + \sigma_{j+1}^2} - \sqrt{\sigma_1^2 + \sigma_J^2} \right]. \quad (\text{B7})$$

We substitute the observed marginal rates $\bar{p}_j$ for the unknown common p when computing each σ_j^2 . The metric M reports the share of net displacement in noise-corrected total movement; trajectories whose observed reversal lies at or below the noise floor are treated as effectively monotonic ($M = 1$), with the threshold $\max(\mathbb{E}[T - N \mid H_0], (J-1)/n_{\text{eff}})$ guarding against false reversal signals at very small effect sizes.

B.3 Power

The sampling-noise corrections are not optional refinements. At $n_a = 100$, Bernoulli sampling contributes a per-cell standard deviation of up to $\sqrt{0.25/100} = 0.05$. Without correction, a perfectly stable model with verdict rates near 0.5 would register as less stable than one with rates near 0 or 1 as a pure sampling artifact, and S would scale with n_a rather than measuring a fixed property of the model deployment. The analogous correction for M ensures that trajectories whose total variation is consistent with constant-rate sampling noise receive $M = 1$ rather than being penalized for jitter that any finite-sample evaluation will display.

For stability, the principal effect size of interest is systematic between-paraphrase variation of $\sigma_{\text{between}} = 0.05$ (corresponding to $S = 1 - 0.05/\sqrt{0.30} \approx 0.91$). At $n_{\text{eff}} = 500$ per escalation level, the design distinguishes $S \approx 0.91$ from $S = 1.0$ at $\alpha = 0.05$ with power exceeding 0.8. For Pareto viability, the aggregated $n_P = 1,000$ disambiguated trials per scenario distinguish a correct-option rate of $c = 0.95$ from $c = 0.99$ (equivalently, $P = 0.90$ from $P = 0.98$) at $\alpha = 0.05$.

References

- Anderson, E. (1995). *Value in ethics and economics*. Cambridge, MA: Harvard University Press.
- Andriushchenko, M., Souly, A., Dziemian, M., Duenas, D., Lin, M., Wang, J., ... Davies, X. (2025). AgentHarm: a benchmark for measuring harmfulness of LLM agents. *Proceedings of the 13th international conference on learning representations (ICLR)*.

- Arendt, H. (1958). *The human condition* (2nd ed.). Chicago: University of Chicago Press.
- Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., ... Kaplan, J. (2021). *A general language assistant as a laboratory for alignment*. (Preprint at <https://arxiv.org/abs/2112.00861>)
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... Kaplan, J. (2022). *Constitutional AI: harmfulness from AI feedback*. (Preprint at <https://arxiv.org/abs/2212.08073>)
- Bean, A.M., Kearns, R.O., Romanou, A., Hafner, F.S., Mayne, H., Batzner, J., ... Mahdi, A. (2025). Measuring what matters: construct validity in large language model benchmarks. *Advances in neural information processing systems, datasets and benchmarks track*.
- Bengio, Y., Clare, S., Prunkl, C., Murray, M., Andriushchenko, M., Bucknall, B., ... Mindermann, S. (2026). *International AI Safety Report 2026* (Tech. Rep. No. DSIT 2026/001). Department for Science, Innovation and Technology. (<https://internationalaisafetyreport.org>)
- Berlin, I. (1969). Two concepts of liberty. *Liberty* (2nd ed., pp. 166–217). Oxford: Oxford University Press.
- Casper, S., Davies, X., Shi, C., Gilbert, T.K., Scheurer, J., Rando, J., ... Hadfield-Menell, D. (2023). *Open problems and fundamental limitations of reinforcement learning from human feedback*. (Preprint at <https://arxiv.org/abs/2307.15217>)
- Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems* (Vol. 30, pp. 4302–4310). Red Hook, NY: Curran Associates.
- Conitzer, V., Freedman, R., Heitzig, J., Holliday, W.H., Jacobs, B.M., Lambert, N., ... Zwicker, W.S. (2024). Position: social choice should guide AI alignment in dealing with diverse human feedback. *Proceedings of the 41st international conference on machine learning (ICML)* (pp. 9346–9360).
- Desai, N., Critch, A., Russell, S.J. (2018). Negotiable reinforcement learning for Pareto optimal sequential decision-making. *Advances in neural information processing systems* (Vol. 31). Red Hook, NY: Curran Associates.
- Dillion, D., Mondal, D., Tandon, N., Gray, K. (2025). AI language model rivals expert ethicist in perceived moral expertise. *Sci. Rep.*, 15(1), 4084, <https://doi.org/10.1038/s41598-025-86510-0>

- Dobbe, R., Krendl Gilbert, T., Mintz, Y. (2021). Hard choices in artificial intelligence. *Artif. Intell.*, 300, 103555, <https://doi.org/10.1016/j.artint.2021.103555>
- Eckersley, P. (2019). *Impossibility and uncertainty theorems in AI value alignment (or why your AGI should not have a utility function)*. (Preprint at <https://arxiv.org/abs/1901.00064>)
- El, B., & Zou, J. (2025). *Moloch’s bargain: emergent misalignment when LLMs compete for audiences*. (Preprint at <https://arxiv.org/abs/2510.06105>)
- Evans, J., Bratton, B., Agüera y Arcas, B. (2026). *Agentic AI and the next intelligence explosion*. (Preprint at <https://arxiv.org/abs/2603.20639>)
- Fazelpour, S., & Fleisher, W. (2025). The value of disagreement in AI design, evaluation, and alignment. *Proceedings of the 2025 ACM conference on fairness, accountability, and transparency (FAccT)* (pp. 2138–2150). New York: Association for Computing Machinery.
- Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds Mach.*, 30(3), 411–437, <https://doi.org/10.1007/s11023-020-09539-2>
- Gabriel, I., & Keeling, G. (2025). A matter of principle? AI alignment as the fair treatment of claims. *Philos. Stud.*, 182(7), 1951–1973, <https://doi.org/10.1007/s11098-025-02300-4>
- Gabriel, I., Manzini, A., Keeling, G., Hendricks, L.A., Rieser, V., Iqbal, H., ... Manyika, J. (2024). *The ethics of advanced AI assistants*. (Preprint at <https://arxiv.org/abs/2404.16244>)
- Gausen, A., Wallbridge, S., O’Dell, B., Summerfield, C., Kirk, H.R. (2026). *RealityTest: how people probe AI identity and whether models disclose it*. (Preprint at <https://arxiv.org/abs/2606.00168>)
- Ge, L., Halpern, D., Micha, E., Procaccia, A.D., Shapira, I., Vorobeychik, Y., Wu, J. (2024). Axioms for AI alignment from human feedback. *Advances in neural information processing systems* (Vol. 37).
- Haas, J., Bridgers, S., Manzini, A., Henke, B., May, J., Levine, S., ... Isaac, W. (2026). A roadmap for evaluating moral competence in large language models. *Nature*, 650(8102), 565–573, <https://doi.org/10.1038/s41586-025-10021-1>
- Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., ... Wellman, M. (2025). *Multi-agent risks from advanced AI* (Tech. Rep.).

- Cooperative AI Foundation. (Preprint at <https://arxiv.org/abs/2502.14143>)
- Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., Steinhardt, J. (2021). Aligning AI with shared human values. *Proceedings of the 9th international conference on learning representations (ICLR)*.
- Henselmans, D.R., Libert, A., Pier van Gosliga, S., Zwart, L. (2026). *Unified AI alignment is impossible*. (Manuscript. <https://philarchive.org/rec/DAAUAA>)
- Henselmans, D.R., Prinzhorn, D.W.E., Libert, A. (2026). Measuring AI accountability through argumentation analysis: can model reasoning withstand scrutiny? *Paris J. AI Digit. Ethics*, ,
- Huang, S., Siddarth, D., Lovitt, L., Liao, T.I., Durmus, E., Tamkin, A., Ganguli, D. (2024). Collective constitutional AI: aligning a language model with public input. *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency (FAccT)* (pp. 1395–1417). New York: Association for Computing Machinery.
- Jiao, J., Afroogh, S., Murali, A., Chen, K., Atkinson, D., Dhurandhar, A. (2025). LLM ethics benchmark: a three-dimensional assessment system for evaluating moral reasoning in large language models. *Sci. Rep.*, 15(1), 34642, <https://doi.org/10.1038/s41598-025-18489-7>
- Kasirzadeh, A. (2024). Plurality of value pluralism and AI value alignment. *Pluralistic alignment workshop at NeurIPS 2024*.
- Kilov, D., Hendy, C., Guyot, S.Y., Snoswell, A.J., Lazar, S. (2025). *Discerning what matters: a multi-dimensional assessment of moral competence in LLMs*. (Preprint at <https://arxiv.org/abs/2506.13082>)
- Krügel, S., Ostermaier, A., Uhl, M. (2023). ChatGPT’s inconsistent moral advice influences users’ judgment. *Sci. Rep.*, 13(1), 4569, <https://doi.org/10.1038/s41598-023-31341-0>
- Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., Legg, S. (2018). *Scalable agent alignment via reward modeling: a research direction*. (Preprint at <https://arxiv.org/abs/1811.07871>)
- Levine, S., Franklin, M., Zhi-Xuan, T., Guyot, S.Y., Haas, J., Wong, L., ... Gabriel, I. (2026). Resource rational contractualism should guide AI alignment. *Proc. IASEAI Conf.*, 2(1), 353–368,

- Lynch, A., Wright, B., Larson, C., Troy, K.K., Ritchie, S.J., Mindermann, S., ... Hubinger, E. (2025). *Agentic misalignment: how LLMs could be an insider threat*. (Anthropic Research. <https://www.anthropic.com/research/agent-misalignment>)
- Mazeika, M., Yin, X., Tamirisa, R., Lim, J., Lee, B.W., Ren, R., ... Hendrycks, D. (2025). *Utility engineering: analyzing and controlling emergent value systems in AIs*. (Preprint at <https://arxiv.org/abs/2502.08640>)
- Mishra, A. (2023). *AI alignment and social choice: fundamental limitations and policy implications*. (SSRN Working Paper 4605509)
- Oh, S., & Demberg, V. (2025). Robustness of large language models in moral judgements. *R. Soc. Open Sci.*, 12(4), 241229, <https://doi.org/10.1098/rsos.241229>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. A.H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in neural information processing systems* (Vol. 35, pp. 27730–27744).
- Pan, A., Chan, J.S., Zou, A., Li, N., Basart, S., Woodside, T., ... Hendrycks, D. (2023). Do the rewards justify the means? Measuring trade-offs between rewards and ethical behavior in the Machiavelli benchmark. *Proceedings of the 40th international conference on machine learning (ICML)* (pp. 26837–26867). Honolulu, HI: PMLR.
- Qiu, T. (2024). Representative social choice: from learning theory to AI alignment. *Pluralistic alignment workshop at NeurIPS 2024*.
- Rudschies, C., Schneider, I., Simon, J. (2020). Value pluralism in the AI ethics debate: different actors, different priorities. *Int. Rev. Inf. Ethics*, 29, , <https://doi.org/10.29173/irie419>
- Russell, S.J. (2019). *Human compatible: artificial intelligence and the problem of control*. London: Penguin Books.
- Scherrer, N., Shi, C., Feder, A., Blei, D. (2023). Evaluating the moral beliefs encoded in LLMs. *Advances in neural information processing systems* (Vol. 36).
- Scheurer, J., Balesni, M., Hobbhahn, M. (2024). *Large language models can strategically deceive their users when put under pressure*. (Preprint at <https://arxiv.org/abs/2311.07590>)

- Sclar, M., Choi, Y., Tsvetkov, Y., Suhr, A. (2024). Quantifying language models’ sensitivity to spurious features in prompt design, or: how I learned to start worrying about prompt formatting. *Proceedings of the 12th international conference on learning representations (ICLR)*.
- Sehwag, U.M., Shabihi, S., McAvoy, A., Sehwag, V., Xu, Y., Towers, D., Huang, F. (2025). *PropensityBench: evaluating latent safety risks in large language models via an agentic approach*. (Preprint at <https://arxiv.org/abs/2511.20703>)
- Sen, A. (1970). The impossibility of a Paretian liberal. *J. Polit. Econ.*, 78(1), 152–157,
- Sharma, M., McCain, M., Douglas, R., Duvenaud, D. (2026). *Who’s in charge? Disempowerment patterns in real-world LLM usage*. (Preprint at <https://arxiv.org/abs/2601.19062>)
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S.R., . . . Perez, E. (2024). Towards understanding sycophancy in language models. *Proceedings of the 12th international conference on learning representations (ICLR)*.
- Siththaranjan, A., Laidlaw, C., Hadfield-Menell, D. (2023). *Distributional preference learning: understanding and accounting for hidden context in RLHF*. (Preprint)
- Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C.M., . . . Choi, Y. (2024). Position: a roadmap to pluralistic alignment. *Proceedings of the 41st international conference on machine learning (ICML)* (pp. 46280–46302). Vienna, Austria: PMLR.
- Tomašev, N., Franklin, M., Jacobs, J., Krier, S., Osindero, S. (2025). *Distributional AGI safety*. (Preprint at <https://arxiv.org/abs/2512.16856>)
- van Nuenen, T., & Sachdeva, P.S. (2026). *The fragility of moral judgment in large language models*. (Preprint at <https://arxiv.org/abs/2603.05651>)
- Wei, A., Haghtalab, N., Steinhardt, J. (2023). Jailbroken: how does LLM safety training fail? *Advances in neural information processing systems* (Vol. 36, pp. 80079–80110). Red Hook, NY: Curran Associates.
- Zhi-Xuan, T., Carroll, M., Franklin, M., Ashton, H. (2025). Beyond preferences in AI alignment. *Philos. Stud.*, 182(7), 1813–1863, <https://doi.org/10.1007/s11098-024-02249-w>