Decision Divergence Among Large Language Models in Organizational Scenarios Without Ground Truth

Rafael Camacho-Aguilar
Escuela de Administración de Empresas
Instituto Tecnológico de Costa Rica
Cartago, Costa Rica
fitacam@gmail.com

Javier Rojas-Segura
Escuela de Administración de Empresas
Instituto Tecnológico de Costa Rica
San José, Costa Rica
jarojas@itcr.ac.cr

José Martínez-Villavicencio
Escuela de Administración de Empresas
Instituto Tecnológico de Costa Rica
Cartago, Costa Rica
jomartinez@itcr.ac.cr

***Abstract*—Organizations increasingly delegate consequential judgments to large language models in settings that have no ground truth, where no label exists against which a recommendation could be scored. Reliability then cannot be framed as accuracy; what remains measurable is whether a different model would have answered differently. We put twelve organizational decision scenarios, crossing three business domains with four consequence levels, to three instruction-tuned models from three independent providers under a full factorial of temperature and prompting strategy (1,080 inferences), and define the Decision Divergence Rate (DDR) as the share of configurations in which the three models fail to agree. Treating the scenario as the unit of inference ($n = 12$), DDR is 0.247. The models differ in how readily they recommend acting (38.2%, 32.9% and 25.0% of cells; Llama–Gemini difference 0.132, cluster-bootstrap 95% CI $[0.006, 0.270]$), a contrast that rests on four scenarios and that we report descriptively, and self-reported confidence separates agreement from disagreement ($d = 0.96$; CI $[0.66, 1.24]$). Divergence did not track the ambiguity we designed into the tasks: it concentrated in tactical trade-offs, and wherever the three models did agree they agreed on the conservative option. We argue that such consensus is evidence about shared alignment pressure, not about the world.**

***Keywords*—large language models, decision divergence, predictive multiplicity, organizational decision-making, AI governance**

I. Introduction

Organizations delegate a growing share of managerial judgment to automated systems. Data-driven decision-making spread quickly across firms and is associated with measurable productivity gains [1], and algorithms in the workplace no longer merely inform: they direct, evaluate and discipline [2]. Part of what drives the delegation is a mismatch that predates these systems: recording capacity has grown far faster than the capacity to process and analyse what is recorded, and the gap is attributed to the speed of data management and storage technologies outpacing human capacity to turn collected data into knowledge [3]. Large language models (LLMs) extend that delegation into territory the earlier generation of systems did not reach. Where a scoring model returned a number, a language model returns a recommendation with a rationale attached, which makes it usable for the kind of open-ended executive question that had remained human [4].

Many of those questions have no answer key. Whether to approve a rollout, authorize a restructuring, or commit to a supplier standard is not a task with a label waiting to be revealed; the counterfactual never becomes observable, and the organization must act before any outcome could settle the matter. The reliability question therefore changes shape. It cannot be whether the model is accurate, because accuracy is undefined in the absence of a criterion. What remains measurable is whether the recommendation would have been the same had the organization deployed a different model.

That reformulation is not merely methodological. If two comparably credible systems return opposite recommendations on the same case, the deployment choice—ordinarily treated as a procurement detail—silently determines the decision, and no appeal to correctness can adjudicate it. This is the condition predictive multiplicity describes for classifiers [5], extended by subsequent work into a question of institutional governance rather than model selection [6]. It acquires additional force for language models, whose outputs already vary with decoding procedure [7] and across repetitions of an identical task [8], and whose fluency offers the reader no signal about grounding [9].

The divergence question, however, has a counterpart that receives far less attention. If independently developed models converge, an organization reads the agreement as corroboration. Yet the welfare consequences of many decision-makers relying on correlated systems are themselves adverse [10], and shared components across deployed systems produce homogenized outcomes for the same individuals [11]. Whether inter-model agreement carries evidential weight in a setting without ground truth is an open question, and one that cannot be settled without first measuring how, where and when these systems disagree.

Existing work approaches that measurement and stops short of it. Multiplicity has been studied within a model class on labeled tasks. Research on LLM-as-judge evaluates models against human preference [12], which restores a criterion

by proxy, and has established that different evaluator models judge the same output differently [13]. Self-consistency measures agreement across samples drawn from a single model [14] rather than between models. None of these addresses the case an organization actually faces: several independently developed models, one decision, and no key.

This paper takes up that case. We put twelve organizational decision scenarios, crossing three business domains with four consequence levels, to three instruction-tuned models from three independent providers, under a full factorial of temperature and prompting strategy. We define the *Decision Divergence Rate* (DDR) as the proportion of decision configurations in which the three models fail to agree, and we treat the scenario as the unit of inference, which makes the study exploratory by construction. We ask four questions:

RQ1 How often do independently developed models diverge on the same organizational decision?

RQ2 Do the models exhibit decision biases that shape which recommendation an organization receives?

RQ3 Can divergence be anticipated from the models' own reported confidence?

RQ4 Where does divergence concentrate, and does it track the ambiguity designed into the task?

We answer these questions with a design that treats the absence of a key as the object of measurement rather than as a defect to be corrected. The remainder of this paper is organized as follows. Section II situates the study within work on multiplicity, monoculture, and evaluation without a criterion. Section III describes the scenario battery, the generation protocol, and the cluster-robust inference procedure. Section IV reports the evidence for RQ1–RQ4. Section V draws out the governance implications and the limitations, and Section VI concludes.

II. Background and Related Work

A. Delegated Judgment and the Limits of the Data-Driven Firm

The adoption of data-driven decision-making was rapid and broad, and it correlates with productivity at the firm level [1]. Adoption, however, is not practice. Firms that describe themselves as data-driven frequently are not, and the distance between the two is mediated by digital orientation and by who inside the organization holds the license to interpret the numbers [15]. What is delegated is therefore never only computation; it is also interpretive authority. Algorithms at work direct, evaluate and discipline, which makes the choice of system a contested question of organizational control rather than a matter of technical efficiency [2].

That delegation meets a recipient who is not a neutral evaluator. The manager supervising an automated recommendation is a bounded reasoner, subject to anchoring, framing and overconfidence before any system is introduced [16]. The volume of available data compounds those limits rather than relieving them. The gap between the data on offer and the capacity to process it has been characterized as *perceived data overload*, a condition associated with technostress and with degraded individual performance [17]. A system that returns a reasoned recommendation rather than another table is therefore attractive for reasons that have little to do with whether its reasoning is sound. Field evidence complicates the assumption that a better model therefore yields a better decision: a superior algorithm can degrade organizational performance when the surrounding human interaction fails to adapt to it [18]. The behavioral response is itself unstable, swinging between algorithm aversion and algorithm appreciation according to task and framing [19]. Generative systems inherit these dynamics and add capabilities the earlier literature did not anticipate [4].

B. Reliability When There Is No Key

Almost everything established about LLM reliability rests on tasks that have a criterion. Benchmarks for table-based fact verification, for instance, score a model against labeled entailment [20]. The difficulty appears as soon as the construct is latent rather than factual: for sentiment, political leaning or sarcasm, annotations by models and by humans diverge, and which of the two should serve as the standard stops being obvious [21].

Three properties of these systems compound that difficulty. Fluency is not grounding, so a model can produce well-formed prose that no evidence supports [9]. Output depends on the decoding procedure as much as on the model, so the same system can answer the same prompt differently [7], and that instability persists at the level of reasoning tasks: models give significantly inconsistent responses across repetitions, and when they err they often err in ways that do not resemble human cognitive biases [8]. Self-reported confidence does not resolve the problem either. Models asked to verbalize uncertainty are systematically overconfident and compress their answers into a narrow band at the top of whatever scale they are offered [22].

C. Multiplicity, Monoculture, and Two Ways of Failing

The framework that fits a regime without a criterion is predictive multiplicity. Models fit to the same data with comparable accuracy assign conflicting predictions to individual cases, which makes the model an institution deploys a policy choice rather than a technical detail [5]. Work on model multiplicity carries that observation into governance and asks who absorbs the cost of an arbitrary selection [6]. For language models the closest empirical analogue is the LLM-as-judge literature: evaluator models are ranked by their agreement with human preference [12], and they are demonstrably not neutral, since different models judge identical outputs differently and in patterned ways [13]. Within a single model, agreement across sampled reasoning paths already serves as a proxy for reliability [14].

Disagreement, however, is only one of the two ways this arrangement can fail, and the literature on algorithmic monoculture describes the other. When many decision-makers converge on a single algorithm, aggregate decision quality can fall even though the algorithm is more accurate for each of

them in isolation [10]. Systems built from shared components extend that concern from welfare to distribution, producing homogenized adverse outcomes for the same individuals across independently operated deployments [11].

There is reason to expect language models to be particularly exposed to this second failure. Alignment by reinforcement learning from human feedback collapses the response distribution rather than preserving its spread: tuned models concentrate on modal positions instead of reflecting the diversity of the populations that produced their feedback data [23]. The procedure carries known and systematic limitations that a single training pipeline cannot resolve, and those limitations are shared across the providers who employ it [24]. Models trained on overlapping web corpora and aligned by comparable procedures therefore have grounds to agree for reasons unrelated to the merits of the case before them.

D. The Gap

These strands converge on our question and each stops short of it. Multiplicity is measured within a model class on labeled data. LLM-as-judge restores a criterion through human preference. Self-consistency measures agreement inside one model rather than between models. Monoculture research anticipates the cost of convergence but addresses institutions adopting one shared system, not several independently developed systems answering the same question. What has not been measured is how far such models diverge from one another on organizational decisions where no key exists—at what rate, where that divergence concentrates, whether the models themselves signal it, and what their agreement is worth when it occurs.

We should be precise about where we claim novelty, because the measure itself does not carry it. The Decision Divergence Rate is the proportion of non-unanimous responses, and as an operation it is elementary; related quantities appear wherever inter-rater agreement is discussed. What we claim is narrower and empirical: that this quantity, applied to organizational decisions that have no key and computed with the scenario rather than the response as the unit of inference, yields three findings not previously reported for this setting—divergence runs *against* designed ambiguity rather than with it, it concentrates in tactical trade-offs, and unanimity when it occurs is one-directional. The contribution is the organizational application and the pattern it reveals, not the arithmetic of the index.

III. Method

A. Study Design

This paper studies how LLMs diverge when they are used for organizational decisions that have no ground truth. Accuracy is not our object of study. In the setting we examine no correct answer exists against which a model could be scored, so a decision-maker cannot settle a disagreement by asking which model is right. What remains available is agreement, and its absence.

This is the regime that predictive multiplicity describes [5], [6]: models that a practitioner has no principled basis for ranking can still assign different outcomes to the same case. We make no claim that the three models we query are equally capable, because our design gives us no way to establish one—and that is deliberate. The inability to adjudicate is the condition we set out to study, not a gap in the design.

We treat the study as exploratory throughout. The scenario is our unit of inference, which leaves $n = 12$—enough to characterize divergence and generate hypotheses, but not to confirm them.

B. Scenario Battery

We built twelve scenarios as a fully crossed 3×4 design: three business domains—talent management, environmental, social and governance (ESG), and marketing—by four consequence levels, with one scenario per cell (Table I). All twelve share the same instrumentation template. Each one presents an organizational context, and then places a three-year historical trend in *directional conflict* with a recent signal—that conflict is our intended generator of ambiguity. Each then reports three KPIs whose favorable and unfavorable ordering we alternate across scenarios, to control for primacy and recency effects, and closes with a binary executive question that opposes acting (approve, authorize, or commit) to holding back (reject or defer). We name no industry, country, or organization, which limits recall of memorized specifics, and we write every scenario in English.

Each scenario also carries two author-assigned values, both recorded in the scenario definition code before we ran any inference: an ambiguity score from 3 to 5, and a prior on the expected divergence rate, `expected_ddr`. We use that prior only as a manipulation check on the ambiguity construct. It was time-stamped in code but not deposited in a public registry, so we call it *pre-specified* rather than pre-registered.

Fig. 1 reproduces one scenario in the form the models received it, abridged for space.

One property of the battery constrains everything that follows. The ambiguity score is not orthogonal to consequence level: as Table I shows, every strategic and every ethical scenario carries the maximum score of 5, so the two factors are collinear here and their effects cannot be separated. We state this in the design rather than in the limitations because it governs how the results of Section IV can be read, and we return to it in Section V.

C. Models and Generation Protocol

We query three instruction-tuned LLMs from three independent providers: Llama 3.3 70B, DeepSeek-V3, and Gemini 3.1 Flash-Lite. We present each scenario to each model under a full factorial of three temperatures (0.0, 0.5, 1.0) and two prompting strategies (zero-shot and chain-of-thought expert), with five iterations per configuration. That gives 30 inferences per model per scenario, and $N = 1080$ in total.

The two strategies differ only in the system instruction and in the required output fields, summarized in Table II. The scenario block itself is identical across strategies, so any difference we observe between them is attributable to the

TABLE I
SCENARIO BATTERY. AMB. IS THE AUTHOR-ASSIGNED AMBIGUITY SCORE; E[DDR] IS THE PRE-SPECIFIED PRIOR.

ID	Domain	Consequence	Amb.	E[DDR]
S01	Talent mgmt.	Operational	4	high
S02	Talent mgmt.	Tactical	3	medium
S03	Talent mgmt.	Strategic	5	very high
S04	Talent mgmt.	Ethical	5	very high
S05	ESG	Operational	3	medium
S06	ESG	Tactical	4	high
S07	ESG	Strategic	5	very high
S08	ESG	Ethical	5	very high
S09	Marketing	Operational	4	high
S10	Marketing	Tactical	4	high
S11	Marketing	Strategic	5	very high
S12	Marketing	Ethical	5	very high

TABLE II
PROMPTING STRATEGIES. THE SCENARIO BLOCK IS IDENTICAL UNDER BOTH; ONLY THE SYSTEM INSTRUCTION AND THE REQUIRED JSON FIELDS DIFFER.

Strategy	System instruction and required output
Zero-shot	Persona: senior business advisor. Analyze the indicators provided and issue a binary recommendation. Output fields: `decision`, `confidence`, `decisive_kpi`, `justification` (3 sentences maximum).
Chain-of-thought expert	Persona: strategic decision auditor. Follow five ordered steps before recommending: (1) how the recent signal contradicts the historical trend; (2) whether each KPI supports or undermines action; (3) reversibility, given that operational decisions can be corrected while strategic and ethical ones set precedents; (4) the single decisive piece of evidence; (5) the binary recommendation. Same output fields plus `reasoning` (4–6 sentences).

S06 • ESG/sustainability • tactical • ambiguity 4 • prior E[DDR]: high

Context. The procurement team proposes implementing a mandatory ESG certification requirement for all 80 Tier-1 suppliers, with an 18-month enforcement timeline. A pre-implementation audit reveals that 40% of the current supplier base—32 of 80 suppliers—would not meet the proposed standard as of today. Of the 32 non-compliant suppliers, 14 provide components for which no qualified alternative source has been identified. Two institutional investors holding a combined 8.3% of outstanding shares have sent a formal letter requesting a concrete supply chain ESG compliance timeline.

Historical trend (last 3 years). The supplier base grew from 51 to 80 Tier-1 suppliers as production volume expanded. Each annual sustainability report has included language committing to "responsible supply chain practices," but no formal ESG supplier screening or certification program has ever been operationalized. [. . .]

Recent signal. The investors' joint letter explicitly references the organization's sustainability report language as "aspirational but unmeasurable." Separately, a major customer representing 11% of revenue has indicated it will require ESG supply chain certification from all vendors as a contract renewal condition starting in 24 months. [. . .]

Key performance indicators.
• Tier-1 suppliers failing proposed ESG threshold: 40% (32 of 80).
• Non-compliant suppliers with no qualified alternative source: 14 of 32 (single-source dependency).
• Estimated supply disruption cost if non-compliant suppliers are terminated on schedule: $3.2M.

Executive question. Do you approve the mandatory ESG supplier certification requirement with 18-month enforcement timeline?

Fig. 1. One scenario as presented to the models (S06, abridged). The directional conflict between the historical trend and the recent signal is the intended generator of ambiguity; the KPIs are ordered so that support for and against action alternate.

reasoning instruction rather than to the stimulus. We constrain every response with a strict JSON schema requiring a binary decision, a self-reported confidence rating on a 1–10 scale, the single indicator the model considered decisive, and a written justification. Of the 1080 responses, 1076 parsed successfully (4 format failures, 0.37%), which leaves 356 of 360 cells with all three model decisions present.

D. Measures

We call a *cell* the triple of model decisions for one (scenario, temperature, strategy, iteration) configuration, and we call a cell *divergent* when the three decisions are not unanimous. The DDR is then the proportion of divergent cells within a grouping. DDR is a property of a decision task under a given model ensemble, not a property of any one model: it measures how often the choice of model would change the recommendation that reaches a decision-maker.

We deliberately do not report chance-corrected agreement coefficients such as Fleiss' κ. With three raters, a binary outcome and no ground truth, κ absorbs differences in the raters' marginal response rates into its chance-agreement term. Those marginal differences are themselves one of our findings (Section IV), so the correction would net out the very effect we want to measure. DDR keeps the two quantities separable.

DDR was flat across the factors we manipulated during generation, varying negligibly by temperature (0.250, 0.244, 0.248) and by prompting strategy (0.257 versus 0.237). We therefore aggregate over both factors throughout.

E. Statistical Inference

Ambiguity score is constant within a scenario, and that makes the scenario, not the cell, our unit of inference; $n = 12$. Treating the 356 cells as independent observations would be pseudoreplication, and the data show why: differences between scenarios account for 71% of the total variance in divergence, an intraclass correlation of 0.71. We therefore regard cell-level p-values as invalid, and where we report one it is only to declare it so.

All our interval estimates come from a cluster bootstrap that resamples entire scenarios with replacement ($B = 10{,}000$, seed 20260717), recomputing each statistic on the resampled battery. For the association between designed ambiguity and observed divergence we use an exact permutation test instead, enumerating all 13,860 distinct assignments of the ambiguity multiset $\{3^2, 4^4, 5^6\}$ to the twelve scenarios, so that no asymptotic approximation enters. Resampling whole scenarios also absorbs within-scenario replication, which matters here because generation at temperature 0 is near-deterministic—only 1 of 72 configurations varied across its five iterations—so those five iterations are close to copies rather than independent draws. Within-model variation remained limited at the

TABLE III
DECISION DIVERGENCE RATE BY DESIGN FACTOR. ALL VALUES ARE DESCRIPTIVE; THE INFERENTIAL RESULTS ARE REPORTED IN THE TEXT.

Grouping	Level	DDR
Overall	—	0.247
Consequence level	Tactical	0.800
	Ethical dilemma	0.144
	Strategic	0.023
	Operational	0.011
Domain	ESG/sustainability	0.381
	Talent management	0.244
	Marketing	0.118
Ambiguity score	3	0.467
	4	0.381
	5	0.084

higher temperatures as well: 5 of 72 configurations varied at temperature 0.5 and 10 of 72 at 1.0.

IV. RESULTS

A. Decision Divergence Rate

We first ask how often the three models disagree at all. Of the 356 complete cells, 88 diverged, which puts the Decision Divergence Rate at 0.247 with a cluster-bootstrap 95% interval of $[0.05, 0.48]$. In roughly one decision configuration in four, the recommendation an organization received depended on which of the three models it had deployed. The interval is wide, as it must be with twelve scenarios, and we read the point estimate as an order of magnitude rather than as a precise rate.

Throughout this section our interval estimates come from the cluster bootstrap of Section III-E. Where we quote a statistic computed at the cell level we say so, because Section III-E treats those p-values as invalid.

Table III breaks the rate down by each design factor. The aggregate conceals more than it reports. Divergence is almost absent in operational and strategic scenarios, at 0.011 and 0.023, and close to universal in tactical ones, at 0.800. We report the domain and ambiguity breakdowns descriptively and attach no test to them. We return to what that concentration means in Section IV-D.

B. Model Decision Bias

Models could disagree without any of them being more or less inclined to act. Descriptively, that is not what we find; inferentially, the battery is too small to say more, and we report the pattern on the same footing as the ambiguity association of Section IV-D: as a descriptive result with its interval, and with no test attached.

The marginal rates give the difference its direction. Across the 356 complete cells Llama recommends acting in 38.2% of them, DeepSeek in 32.9%, and Gemini in 25.0% (Table IV). The two extremes are nearly one-directional: Gemini and Llama disagree in 49 cells, and in 48 of those it is Llama choosing to act where Gemini declines. The cluster-bootstrap 95% interval for the Llama–Gemini difference in action rate is $[0.006, 0.270]$ around a point estimate of 0.132; the intervals for DeepSeek–Gemini (0.079, $[-0.034, 0.203]$) and Llama–DeepSeek (0.053, $[-0.077, 0.206]$) include zero.

TABLE IV
MODEL ACTION RATES AND DISSENT. ACTION RATE IS THE SHARE OF THE 356 COMPLETE CELLS IN WHICH THE MODEL RECOMMENDED ACTING. MINORITY VOTE COUNTS THE DIVERGENT CELLS IN WHICH THE MODEL WAS THE LONE DISSENTER.

Model	Action rate (%)	Minority vote (of 88)
Llama 3.3 70B	38.2	26
DeepSeek-V3	32.9	39
Gemini 3.1 Flash-Lite	25.0	23

Those intervals are wide for a reason that the scenario-level view makes explicit. In six of the twelve scenarios the three models return identical action rates—the five unanimous rejections and the unanimous approval in S05—so the entire Llama–Gemini difference rests on four scenarios (S01, S02, S06, S10), three of them tactical. A thirteen-point spread in action rate is not a rounding difference between interchangeable tools, but with four informative scenarios it is a pattern to be characterized, not a parameter to be tested, and we treat it as such.

What the marginals do not tell us is which model breaks ranks in any given case, and there the answer is less tidy than the aggregate suggests. Across the 88 divergent cells the dissenting vote belongs to DeepSeek 39 times, to Llama 26, and to Gemini 23 (Table IV). No model holds the outlier role; it rotates with the scenario. The sharpest case is S08, the ethical dilemma in the ESG domain, where the dissenter never changes: DeepSeek recommends acting in 13 of its 30 cells while Gemini and Llama recommend it in none.

Taken together the two results point the same way. Models differ in how readily they act, and the identity of the dissenter is a property of the decision rather than of the model. An organization cannot screen once for the cautious model and be done with it.

We do not report Cochran's Q for this comparison, and we flag its absence deliberately. Computed over the 356 cells, Q is pseudoreplicated for the reason Section III-E gives. A cluster-bootstrap interval around an *observed* Q does not fix that: the bootstrap distribution is centred on the sample estimate, not on the null, so comparing its lower bound against a $\chi^2(2)$ critical value tests nothing. The interval on the action-rate difference reported above is the cluster-robust quantity the design supports.

C. Confidence as a Divergence Signal

If divergence cannot be anticipated from the task—and Section IV-D finds no evidence that it can be—then the remaining question is whether the models signal it themselves. We compared self-reported confidence in the responses belonging to consensus cells against those belonging to divergent cells. Mean confidence is 8.47 across the 804 consensus responses and 7.95 across the 264 divergent ones. The standardized difference is $d = 0.96$, with a cluster-bootstrap 95% interval

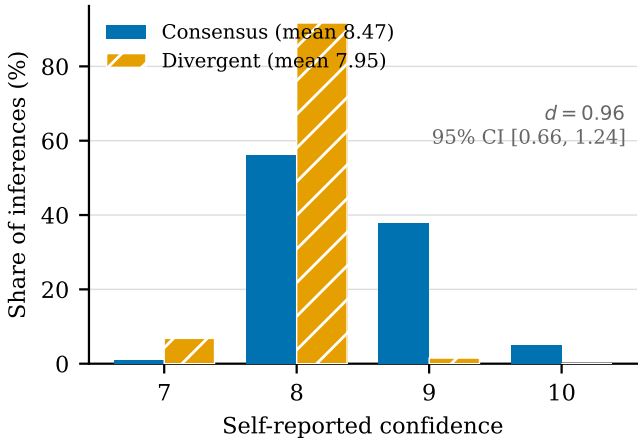


Fig. 2. Self-reported confidence in consensus versus divergent cells. The scale is compressed into the range 7–10; the difference lies in the upper tail.

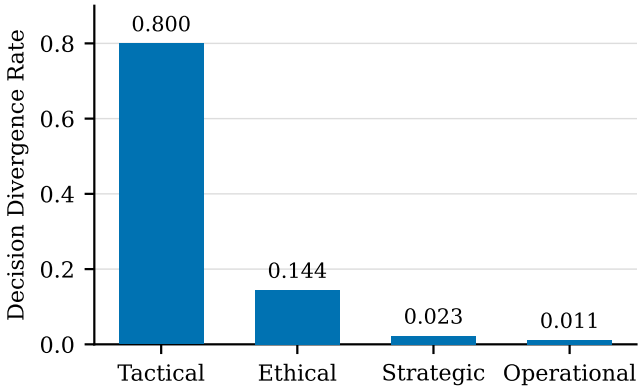


Fig. 3. Decision Divergence Rate by consequence level. Disagreement is concentrated in tactical trade-offs.

of $[0.66, 1.24]$, and no bootstrap replicate produced a value at or below zero. The Welch t for the same comparison is 18.95, computed at the cell level.

Two features of the measurement make that effect more notable rather than less. First, the confidence scale collapsed: models used only the range 7 to 10, and 65% of all responses were exactly 8, which is consistent with the verbalized overconfidence reported for LLMs [22]. A half-point gap on a four-point effective range is a wide one. Second, the effect survives the most conservative correction available to us, which resamples whole scenarios rather than individual responses.

Fig. 2 shows where the two distributions part. Both pile up at 8, but the upper tail belongs almost entirely to consensus: 42.9% of consensus responses exceed a confidence of 8, against 1.5% of divergent ones. The signal is therefore usable in aggregate rather than decisive case by case, and we take up what that permits in Section V.

D. Where Divergence Concentrates

We built the battery on the premise that designed ambiguity produces divergence, and we recorded that premise as a per-scenario prior before running anything. The premise did not hold. Of the six scenarios we marked as very high, five converged: S03, S04, S11 and S12 produced no divergence at all, and S07 produced a DDR of 0.071.

At the cell level, divergence and designed ambiguity correlate at $\rho = -0.38$, with the sign running opposite to what we expected. That coefficient carries a p-value below 10^{-12}, which Section III-E treats as invalid. At the scenario level, which is our unit of inference, $\rho = -0.36$, with an exact permutation p of 0.248 over all 13,860 assignments and a cluster-bootstrap 95% interval of $[-0.85, +0.26]$ (Fig. 4). The interval crosses zero. We report the association as a descriptive pattern and make no inferential claim about it.

What the data do tell us is where divergence lives. Three scenarios—S02, S06 and S10, all of them tactical—account for 72 of the 88 divergent cells, or 82% of all disagreement. The concentration is visible both by consequence level (Fig. 3) and scenario by scenario (Fig. 4). Because every tactical scenario in this battery carries a low or middling ambiguity score while every strategic and ethical one carries the maximum, the statement that divergence concentrates in tactical decisions and the statement that divergence runs against designed ambiguity describe the same three points. Section III-B set out why we cannot separate them.

The converging scenarios turn out to be more informative than the diverging ones. In all six scenarios where the three models agreed unanimously, they agreed on the conservative option: they rejected the risky action in S03, S04, S09, S11 and S12, and approved the compliance upgrade in S05. Unanimity is not distributed across the answer space but points consistently one way.

Two statements must be kept apart here, because conflating them overstates the result. That *unanimity, when it occurs, is conservative* holds in all six cases. That *high-stakes scenarios produce unanimity* does not: of the six strategic and ethical scenarios, S07 and S08 diverged, and S08 did so in 43% of its cells. We make the weaker claim only. It is consistent with a shared risk-averse default rather than with three models independently arriving at the same judgment, and it is the observation we build on in Section V.

V. Discussion

A. Divergence Without a Key, and Its Twin

Our results describe a decision environment that is unstable in a specific way. In a quarter of the decision configurations we put to three models, the recommendation depended on which of them an organization had deployed, and nothing in the task as we designed it anticipated where that would happen. We built the battery expecting ambiguity to drive disagreement, and the scenarios we marked as most ambiguous turned out to be the ones the models agreed on most.

The instability is not spread evenly. It concentrates in tactical trade-offs, where each option carries a defensible case

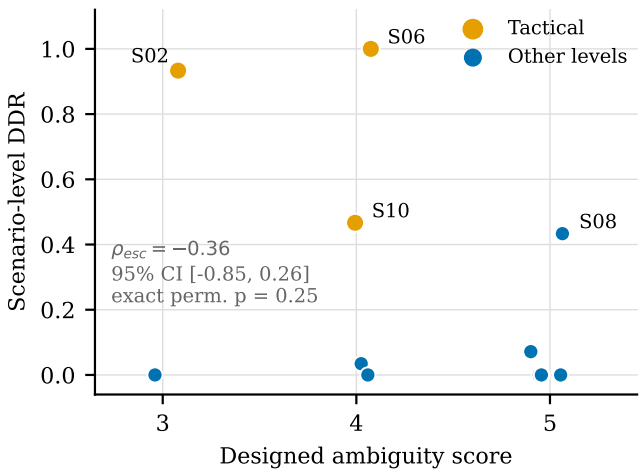


Fig. 4. Scenario-level DDR against designed ambiguity score. The association is descriptive: its cluster-bootstrap interval crosses zero and the exact permutation test does not reject.

and neither is obviously the safe one. Where one option is visibly cautious—halting a rollout, declining a layoff, staying out of a regulatory gray zone—the models converge.

That convergence is the result we did not expect, and we consider it the more consequential of the two. In all six scenarios where the three models agreed unanimously, they agreed on the conservative option. Three models from three independent providers, trained on different corpora under comparable alignment procedures, produced the same default. We read this as consistent with shared alignment pressure rather than with three independent judgments happening to coincide. The mechanism is documented outside our data: alignment by human feedback concentrates a model's response distribution on modal positions instead of preserving the spread of the population that produced the feedback [23], and the limitations of that procedure are structural rather than particular to one provider [24]. We nonetheless acknowledge that our design cannot separate a shared prior from genuine agreement, and we present the reading as an interpretation rather than as a result.

If that reading holds, the risk an organization faces is symmetric, and only one half of it is currently named. The first half is the one predictive multiplicity describes: models disagree, no key exists to adjudicate, and the deployment choice quietly determines the outcome [5], [6]. The second half is its mirror image: models agree, no key exists to adjudicate, and the agreement gets read as corroboration. This is the failure that the algorithmic monoculture literature anticipates from the opposite direction. Convergence of many decision-makers on correlated systems lowers aggregate decision quality even when each system is individually accurate [10], and shared components across deployments concentrate adverse outcomes on the same subjects [11]. Our contribution to that line is to observe the mechanism inside a single decision rather than across a market: inter-model consensus is evidence about the models' shared training pressures, not about the world. An organization that queries three systems and receives one answer has not triangulated; it may have sampled a single prior three times.

The practical asymmetry is that the first risk announces itself and the second does not. Disagreement is visible and forces a human to decide. Agreement is quiet, and in our battery it covers four of the six strategic and ethical scenarios—S03, S04, S11 and S12, the decisions an organization can least afford to get wrong. It does not cover all six, and the two exceptions are worth naming rather than smoothing over: S07 diverged at 0.071 and S08 at 0.433, the latter driven entirely by one model (Section IV-B). The pattern we are describing is therefore that unanimity, when it occurs, is one-directional—not that high-stakes scenarios always produce unanimity.

B. Model Choice as a Governance Decision

The action rates we observe differ by thirteen points between the extremes, in a direction that holds in 48 of the 49 cells where those two models disagree. We do not claim that this difference is established beyond the battery—with four informative scenarios it cannot be—but its direction and its consistency are enough to make the deployment decision a standing policy rather than a technical preference. An organization that standardizes on the most action-prone of the three models will, across a year of decisions, approve materially more than one that standardizes on the most cautious, and no accuracy argument is available to justify either choice.

What our data add is that the policy cannot be reduced to a model's disposition. The dissenting vote rotated across all three models, and which one broke ranks depended on the scenario. A procurement process cannot screen once for the conservative model and consider the matter settled. That places the decision in the territory the ethics-of-algorithms literature maps [25] and that work on model multiplicity assigns to institutions rather than to engineers [6].

C. What Confidence Permits, and What It Does Not

Self-reported confidence separates consensus from divergence with a large effect that survives our most conservative correction, which suggests an operational use: route low-confidence decisions to human review. We want to state the boundary conditions carefully, because the literature does not license treating that handoff as free.

A systematic review and meta-analysis of human–AI combinations finds that on average the combination performs worse than the better of the two acting alone [26]. Human supervisors of automated systems grow complacent in proportion to how reliable the system usually is [27]. Deliberate friction in the interface reduces overreliance, but it costs time and is unpopular with the people subjected to it [28]. Calibrating when to defer remains one of the standing challenges of AI-assisted decision-making [29].

Two considerations bound how far the routing argument can be taken. The first is that human review is not a return to a single correct reading. Eye-tracking evidence on data

interpretation indicates that an insight is produced when a human agent cognitively processes information rather than by its presentation alone, and that different individuals derive divergent interpretations from identical stimuli according to their prior knowledge and biases [30]. Interpretive multiplicity, in other words, is not an artefact of these models; what human review adds is not a criterion but the possibility of deliberation, which is precisely what reading three concurring outputs as corroboration removes.

The second limit comes from our own measurement. The signal is aggregate: divergent cells almost never carry a confidence above 8, yet the large majority of both groups sits at exactly 8. A threshold set at 8 would flag most divergent decisions and most agreeing ones along with them. Confidence is therefore usable as a screening prior over classes of decision rather than as a per-case alarm. The honest formulation is that low reported confidence is grounds to suspect that a second model would have answered differently. It is not grounds to conclude that the answer is wrong.

D. Limitations

Eight limitations bound what we claim. The study rests on twelve scenarios, which is our unit of inference and the reason we present it as exploratory throughout. Ambiguity and consequence level are collinear in this battery, so their effects cannot be separated (Section III-B). The ambiguity score was assigned by the authors and never validated by an external panel, and the falsification of our priors is itself evidence that it failed in its predictive role. The semantics of acting are not uniform: in eleven scenarios acting means accepting risk, while in one it means adopting a compliance measure. The convergence pattern is coherent once read as choosing the safe option, but a cleaner design would balance the direction of action. Generation at temperature 0 is near-deterministic, so the five iterations there are close to copies; resampling whole scenarios absorbs this rather than correcting it directly. The confidence scale compressed into four values, which limits its resolution. There is no ground truth, by design, so we cannot say whether any model was right. Finally, the cell-level p-values our earlier analysis produced were pseudoreplicated, and we report them only to mark them invalid.

VI. Conclusion

Dissent among LLMs on organizational decisions is not anticipated by the property we designed to produce it, but the models signal it themselves. Across twelve scenarios and 1,080 inferences, roughly one decision configuration in four returned a different recommendation depending on which of three models was asked. The three models differ in how readily they recommend acting, and the model that dissents rotates with the decision, which makes the deployment choice a governance decision that no accuracy argument can settle. Where they agreed unanimously, they agreed on the conservative option every time, so inter-model consensus should not be read as corroboration.

Our design leaves clear work ahead. A second study should cross ambiguity and consequence orthogonally so their effects can be separated, validate the ambiguity construct with a panel of human judges rather than by author assignment, balance the direction of the action across scenarios, and elicit confidence on a probabilistic scale wide enough to resolve differences. Because divergence is a property of the models available at a given moment, the battery is also worth re-running periodically against new model releases, treating this run as a baseline at t_0 and homogenization as a quantity that can rise or fall over time. Finally, running the same instrument on tasks that do have a key would establish whether the confidence signal we report behaves the same way when correctness is measurable.

Data Availability

The scenario texts, prompts, raw model outputs and analysis code are available from the authors upon reasonable request.

Declaration on Generative AI

Generative AI tools were used to assist with language editing, structural review of the manuscript, and verification of the statistical code. All methodological decisions, the interpretation of results, and the final wording are the authors' own, and the authors take full responsibility for the content.

References

[1] E. Brynjolfsson and K. McElheran, "The rapid adoption of data-driven decision-making," *American Economic Review*, vol. 106, no. 5, pp. 133–139, 2016.

[2] K. C. Kellogg, M. A. Valentine, and A. Christin, "Algorithms at work: The new contested terrain of control," *Academy of Management Annals*, vol. 14, no. 1, pp. 366–410, 2020.

[3] R. Camacho-Aguilar, J. Rojas-Segura, M. Martínez-Mora, and J. Martínez-Villavicencio, "Applied neuroscience for data visualization," in *Proc. 4th LACCEI Int. Multiconference on Entrepreneurship, Innovation and Regional Development (LEIRD)*, 2024.

[4] M. Choudhury, Z. Elyoseph, N. Fast, D. Ong, E. Nsoesie, and E. Pavlick, "The promise and pitfalls of generative AI," *Nature Reviews Psychology*, vol. 4, no. 2, pp. 75–80, 2025.

[5] C. T. Marx, F. P. Calmon, and B. Ustun, "Predictive multiplicity in classification," in *Proc. 37th Int. Conf. on Machine Learning (ICML)*, ser. PMLR, vol. 119, 2020, pp. 6765–6774.

[6] E. Black, M. Raghavan, and S. Barocas, "Model multiplicity: Opportunities, concerns, and solutions," in *Proc. 2022 ACM Conf. on Fairness, Accountability, and Transparency (FAccT '22)*, 2022, pp. 850–863.

[7] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *Int. Conf. on Learning Representations (ICLR)*, 2020, arXiv:1904.09751.

[8] O. Macmillan-Scott and M. Musolesi, "(Ir)rationality and cognitive biases in large language models," *Royal Society Open Science*, vol. 11, no. 6, 2024.

[9] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proc. 2021 ACM Conf. on Fairness, Accountability, and Transparency (FAccT '21)*, 2021, pp. 610–623.

[10] J. Kleinberg and M. Raghavan, "Algorithmic monoculture and social welfare," *Proceedings of the National Academy of Sciences*, vol. 118, no. 22, 2021.

[11] R. Bommasani, K. A. Creel, A. Kumar, D. Jurafsky, and P. Liang, "Picking on the same person: Does algorithmic monoculture lead to outcome homogenization?" in *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022, arXiv:2211.13972.

[12] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-judge with MT-bench and chatbot arena," in *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*, 2023, pp. 46 595–46 623, arXiv:2306.05685.

[13] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, L. Kong, Q. Liu, T. Liu, and Z. Sui, "Large language models are not fair evaluators," in *Proc. 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Vol. 1: Long Papers*, 2024, pp. 9440–9450.

[14] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," in *Int. Conf. on Learning Representations (ICLR)*, 2023, arXiv:2203.11171.

[15] A. Szukits, "The illusion of data-driven decision making – the mediating effect of digital orientation and controllers' added value in explaining organizational implications of advanced analytics," *Journal of Management Control*, vol. 33, no. 3, pp. 403–446, 2022.

[16] M. H. Bazerman and D. A. Moore, *Judgment in Managerial Decision Making*, 8th ed. John Wiley & Sons, 2012.

[17] B. Giudice da Silva Cezar and A. C. G. Maçada, "Data literacy and the cognitive challenges of a data-rich business environment: An analysis of perceived data overload, technostress and their relationship to individual performance," *Aslib Journal of Information Management*, vol. 73, no. 5, pp. 618–638, 2021.

[18] S. Krakowski, D. Haftor, J. Luger, N. Pashkevich, and S. Raisch, "Human-centered artificial intelligence: A field experiment," *Management Science*, vol. 72, no. 1, pp. 57–72, 2026.

[19] E. Jussupow, I. Benbasat, and A. Heinzl, "An integrative perspective on algorithm aversion and appreciation in decision-making," *MIS Quarterly*, vol. 48, no. 4, pp. 1575–1590, 2024.

[20] W. Chen, H. Wang, J. Chen, Y. Zhang, H. Wang, S. Li, X. Zhou, and W. Y. Wang, "TabFact: A large-scale dataset for table-based fact verification," in *Int. Conf. on Learning Representations (ICLR)*, 2020, arXiv:1909.02164.

[21] L. Bojić, O. Zagovora, A. Zelenkauskaite, V. Vuković, M. Čabarkapa, S. Veseljević Jerković, and A. Jovančević, "Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm," *Scientific Reports*, vol. 15, p. 11477, 2025.

[22] M. Xiong, Z. Hu, X. Lu, Y. Li, J. Fu, J. He, and B. Hooi, "Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs," in *Int. Conf. on Learning Representations (ICLR)*, 2024, arXiv:2306.13063.

[23] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto, "Whose opinions do language models reflect?" in *Proc. 40th Int. Conf. on Machine Learning (ICML)*, ser. PMLR, vol. 202, 2023.

[24] S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak *et al.*, "Open problems and fundamental limitations of reinforcement learning from human feedback," *Transactions on Machine Learning Research*, 2023, arXiv:2307.15217.

[25] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," *Big Data & Society*, vol. 3, no. 2, 2016.

[26] M. Vaccaro, A. Almaatouq, and T. Malone, "When combinations of humans and AI are useful: A systematic review and meta-analysis," *Nature Human Behaviour*, vol. 8, no. 12, pp. 2293–2303, 2024.

[27] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Human Factors*, vol. 52, no. 3, pp. 381–410, 2010.

[28] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 188:1–188:21, 2021.

[29] M. Steyvers and A. Kumar, "Three challenges for AI-assisted decision-making," *Perspectives on Psychological Science*, vol. 19, no. 5, pp. 722–734, 2024.

[30] R. Camacho-Aguilar, J. Rojas-Segura, J. Martínez-Villavicencio, and M. Martínez-Mora, "Seeing an insight: An eye-tracking analysis of how data visualization shapes decision-making," in *Proc. 5th LACCEI Int. Multiconference on Entrepreneurship, Innovation and Regional Development (LEIRD)*, 2025.