Whose Alignment? Comparing LLM Process Alignment Across Diverse Organizational Decision Contexts

Niklas Weller [1] **Emilio Barkett** [2]

Abstract

Aligning AI systems with organizational decision-making is typically framed as a single-target problem: make the model behave like *the* organization. We argue this framing obscures a deeper pluralistic challenge. We rely on a decision-policy capturing method to measure *process alignment*: whether an LLM weights information as the organization does, not merely whether it reaches the same conclusions. Applying this method to ECHR Article 6 decisions, process alignment strongly predicts output accuracy ($r = 0.85$, $p < .001$) and externalization substantially improves alignment for poorly-aligned models. Applying it to German consumer credit decisions, this relationship collapses ($r = 0.15$, $p = .60$): interventions produce inconsistent effects and the benchmark encodes potentially discriminatory historical patterns. This contrast is itself a pluralistic alignment finding: in contested domains, high process alignment is neither achievable via externalization nor unconditionally desirable. Output agreement alone cannot distinguish a model that has internalized an organizational policy from one that merely approximates its outcomes—process-level measurement is a necessary component of any pluralistic alignment evaluation.

1. Introduction

Pluralistic alignment asks how AI systems can faithfully serve stakeholders with diverse, conflicting values rather than converging on a single dominant worldview (Sorensen et al., 2024). Most work in this space focuses on *between-person* plurality: aggregating individual preferences, handling annotation disagreements, or enabling models to adopt different user perspectives. We argue that *between-organization* plurality is an equally important but largely neglected dimension. Organizations are distinct value-encoding institutions: they encode historically contingent decision policies that differ systematically across institutions, domains, and time. Models vary in which organizational value systems they implicitly align with, and this variation is not captured by any existing pluralistic alignment evaluation framework. Recognizing between-organization plurality expands the scope of pluralistic alignment from *whose individual preferences?* to *whose institutional norms?* This question has direct consequences for how AI systems are deployed, audited, and governed in high-stakes settings.

Organizations are repositories of knowledge accumulated through experience, culture, and institutional history (Nonaka, 1994). Some of this knowledge is explicit: rules, policies, codified procedures. Much is tacit: embedded in practice and professional judgment, resistant to full articulation (Polanyi, 1966; Pakarinen & Huising, 2023). When organizations deploy LLMs to decide in line with their intentions and values, they implicitly assume they can transfer this knowledge to a model. The tacit dimension resists this transfer.

This creates a problem distinct from output accuracy. An LLM can reach the right decision for the wrong reasons, weighting information differently than the organization while still achieving adequate accuracy on the observed distribution. We call this divergence the *knowledge gap*: a mismatch between organizational and LLM cue-utilization policies that output agreement fails to detect (Geirhos et al., 2020; Lebovitz et al., 2021). The knowledge gap matters for process legitimacy (Tyler, 1988), generalization to novel cases, and the trustworthiness of AI-assisted decisions in regulated domains.

To operationalize and measure decision process divergence, we rely on a methodology inspired by the Brunswik Lens Model (Brunswik, 1952; Hammond et al., 1964). This allows as to estimate and compare the cue-weighting policy as a proxy for the decision process of both organizations and LLMs from observed decisions. We then computes cosine similarity between these policy vectors as the primary alignment metric. In the following, we refer to this method

[*]Equal contribution [1]University of St. Gallen, Switzerland [2]Columbia University, New York, NY USA. Correspondence to: Niklas Weller <niklas.weller@unisg.ch>, Emilio Barkett <eab2291@columbia.edu>.

Preprint. May 24, 2026.

as the *Contextualized Alignment Lens Model* (CALM) and apply it in two domains:

Study 1 (ECHR): We analyze 1,000 ECHR Article 6 decisions across ten LLM models under three prompting conditions. We find sharp variation in baseline process alignment and show that explicit knowledge externalization substantially improves alignment for poorly-aligned models.

Study 2 (German Credit): We replicate CALM in consumer credit using the German Credit Dataset (Hofmann, 1994). Unlike the legal domain, the credit benchmark is derived from historical decisions that encode potentially discriminatory patterns. We find that alignment and accuracy are orthogonal, that externalization fails to reliably improve alignment, and that at least one model catastrophically over-corrects to base-rate feedback. This is a distinctively pluralistic challenge: *whose values* should the model align to?

Together, these studies demonstrate a core thesis: CALM measures whether an LLM aligns with an organizational decision policy, but whether higher alignment is actually desirable depends on whether that policy itself encodes legitimate values. This is a distinctively pluralistic challenge that output-level evaluation cannot surface. Both studies also raise a third open question: whether a model's stated reasoning faithfully reflects its behavioral policy. CALM measures behavioral policy from observed decisions; whether a model's chain-of-thought tracks or diverges from that inferred policy remains an open challenge we address in the future work section.

2. Background

2.1. The Brunswik Lens Model

The lens model (Brunswik, 1952; Hammond et al., 1964) represents judgment as a linear combination of observable cues: $J = \beta_0 + \sum_i \beta_i C_i$, where C_i are cues and β_i are learned weights. Originally developed for human decision-makers, it measures ecological validity (do cues genuinely predict outcomes?) a achievement (do the decision-maker's weights match ecological realities?). Applied to organizations and LLMs, it reveals whether they weight the same cues in the same direction and magnitude; that is, whether their *observed behavioral policies* are aligned. Meta-analytic evidence confirms that human judges systematically misweight cues relative to ecological validities (Karelaia & Hogarth, 2008); the same divergence between stated and behavioral weights appears when comparing organizational guidelines to individual practitioners' actual decisions (Smith et al., 2003).

We estimate behavioral policies using ridge-regularized logistic regression on observed decisions, fitting identical cue sets to both the organizational benchmark and each LLM. The resulting coefficient vectors $\boldsymbol{\beta}$ represent each decision-maker's inferred cue-utilization policy. Alignment is measured as:

$$\cos(\theta) = \frac{\boldsymbol{\beta}_{\text{org}} \cdot \boldsymbol{\beta}_{\text{LLM}}}{\|\boldsymbol{\beta}_{\text{org}}\| \|\boldsymbol{\beta}_{\text{LLM}}\|} \tag{1}$$

This metric is scale-invariant, symmetric, and bounded in $[-1, 1]$, where 1.0 indicates perfect alignment, 0 orthogonality, and negative values opposition.

2.2. Pluralistic Alignment

Sorensen et al. (2024) distinguish three modes of pluralistic alignment: Overton pluralism (presenting a range of reasonable responses), distributional pluralism (reflecting population-level value distributions), and steerable pluralism (faithfully adopting a specified perspective when directed). The dominant approach to alignment via reinforcement learning from human feedback (Ouyang et al., 2022) tends toward distributional consensus, encoding an aggregate value system rather than supporting genuine institutional plurality (Gabriel, 2020). By making observed organizational cue weightings explicit ("externalization" of organizational knowledge) we use CALM to study the degree of steerable pluralism at the organizational level. This framing requires that steering be *faithful*: the model must genuinely adopt the target policy, not merely mimic its surface outputs. CALM operationalizes this criterion by measuring whether the model's cue-weighting policy shifts toward the target organization's policy under externalization.

Prior work has shown that LLMs can solve judgment tasks at human-level accuracy while weighting information differently (Binz & Schulz, 2023), and that output accuracy alone is insufficient for evaluating AI alignment in organizational settings (Wieringa, 2021; Lebovitz et al., 2021). Our empirical results reinforce this: in one domain, models with very different process alignment scores achieve similar output accuracy, meaning that output-level evaluation can obscures which organizational values a model is actually implementing. Process-level measurement is therefore a necessary component of pluralistic alignment evaluation, not merely a supplementary one, particularly in high-stakes organizational deployment contexts.

3. The CALM Framework

CALM measures process alignment between an organizational benchmark and LLM decision-making across three conditions:

Baseline: The model receives a structured case profile and decides without additional guidance. This reveals the model's *implicit* cue-utilization policy, reflecting what it has

learned through pretraining.

Org-externalized: The model receives the organization's cue-weighting policy, derived from ridge logistic regression on the organization's historical decisions, as explicit guidance before the case. High-weighted cues are described as strong indicators; medium-weighted cues as moderate; low-weighted cues as weak. This tests whether making the tacit explicit closes the knowledge gap.

Introspective-externalized: The model is shown how its own baseline cue-weighting policy diverges from the organization's (aggregate patterns across all cases) and instructed to self-correct. This tests whether self-directed adjustment can supplement or replace explicit guidance.

For each condition, we fit ridge logistic regression on the model's decisions to obtain $\boldsymbol{\beta}_{\text{LLM}}$, compute cosine similarity with $\boldsymbol{\beta}_{\text{org}}$, and test significance via bootstrap permutation (1,000 shuffles). Secondary metrics include Pearson correlation of coefficient vectors, propensity correlation (per-case $P(\text{positive})$ correlation), output accuracy, Cohen's κ, and ROC AUC.

CALM thus serves two functions depending on the domain. Where the organizational benchmark reflects a legitimate, well-specified normative policy, CALM is a *calibration* tool: it quantifies policy divergence and guides externalization to close the gap. Where the benchmark encodes historically contested practices—decision policies that conflict with contemporary legal or ethical standards—CALM is an *audit* tool: it renders the behavioral policy legible so that its normative implications can be examined rather than assumed. The two studies below are designed as deliberate tests of each function.

4. Study 1: ECHR Article 6 Decisions

4.1. Context and Data

We analyze 1,000 ECHR Article 6 decisions (50% violation / 50% no violation) . Cues were coded as 45 binary features organized into nine families (Delay, Counsel, EvidenceAndArms, TribunalIntegrity, etc.) using GPT-5.4-mini with a codebook derived from Article 6 jurisprudence. We tested ten models selected to span sizes, providers, and geographical origin: GPT-5.4, GPT-5.4-mini, GPT-5.4-nano, Grok 4.1 Fast, Claude Haiku 4.5, Gemma 4 (31B), Minimax M2.7, Mistral Large, DeepSeek-v3.2, and Qwen 3.6 Plus, spanning US-, European-, and China-based providers.

The court's ridge logistic regression achieves ROC AUC = 0.714, confirming that binary cues capture meaningful but partial variance in judicial reasoning. The most heavily weighted cue families and overall predictive performance is consistent with prior computational studies of ECHR outcomes (Aletras et al., 2016; Chalkidis et al., 2021).

4.2. Baseline Process Alignment

Baseline process alignment varies widely across the ten models. Cosine similarity ranges from -0.211 (GPT-5.4-nano) to $+0.844$ (GPT-5.4-mini). Three models cluster above $r_{\cos} = 0.82$: GPT-5.4-mini (0.844), Grok 4.1 Fast (0.842), and GPT-5.4 (0.824). Three others show moderate positive alignment: Qwen 3.6 Plus (0.750), Gemma 4 (31B) (0.710), and Minimax M2.7 (0.592). Four models fall near zero or below: Mistral Large (0.083), DeepSeek-v3.2 (0.062), Claude Haiku 4.5 (-0.057), and GPT-5.4-nano (-0.211). This spread does not track general-purpose benchmark performance or pricing tier.

4.3. Effect of Externalization

Org-externalization moves the cosine estimate in the predicted direction for 8 of 10 models, producing statistically significant gains (one-sided $p < .05$) for three initially misaligned models: GPT-5.4-nano ($\Delta = +0.906$), Claude Haiku 4.5 ($\Delta = +0.682$), and Minimax M2.7 ($\Delta = +0.176$). Ceiling effects appear for models already well-aligned: GPT-5.4-mini ($\Delta = +0.045$, n.s.) and GPT-5.4 ($\Delta = +0.045$, n.s.), consistent with the hypothesis that externalization has little room to improve a model whose implicit policy already matches the organization's. Introspective externalization moves the point estimate in the predicted direction for 6 of 10 models; comparable gains appear for Claude Haiku 4.5 ($\Delta = +0.755$), but Grok 4.1 Fast regresses sharply ($\Delta = -0.346$, $p = .002$ in the worsening direction), its well-calibrated implicit policy disrupted by self-correction feedback.

5. Study 2: German Credit Decisions

5.1. Context and the Fairness Tension

The German Credit Dataset (Hofmann, 1994) contains 1,000 consumer credit decisions by a German bank, with a 70% Good / 30% Bad base rate and 20 cues including numerical (loan duration, credit amount, age) and categorical (credit history, employment status, housing) attributes.

This domain introduces a tension absent from the ECHR study. Several cues are protected attributes under contemporary anti-discrimination law: `personal_status_sex`, `age_years`, and `foreign_worker`. The organizational normative model is derived from historical human decisions that may encode discriminatory patterns. The original credit officers may have weighted these attributes in ways that disadvantaged women, older applicants, or foreign workers. High process alignment with this benchmark is therefore not unconditionally desirable (Selbst et al., 2019; Dolata et al., 2022). A model achieving $r_{\cos} \approx 1.0$ with the normative model may be faithfully reproducing institutional discrimination.

This surfaces a core pluralistic alignment question: whose values should the model align to? The organization's historically documented decision policy? Its stated ethical commitments? Contemporary regulatory standards? These targets may conflict, and CALM makes this conflict visible precisely because it measures the *actual* weighting policy rather than stated intent.

5.2. Setup

We apply five LLMs (Claude Haiku 4.5, GPT-5.4-mini, GPT-5.4-nano, Grok 4.1 Fast, DeepSeek-v3.2) under three conditions to all 1,000 cases. These five were selected to span the range of baseline alignment observed in Study 1, from strongly positive ($r_{\cos} > 0.5$) to negative, while limiting the computational cost of running all three conditions across 1,000 cases in this replication study; the remaining Study 1 models will be included in a full replication. To address the original dataset's class imbalance (70/30), we evaluate on a balanced subset of 600 cases (300 Good, 300 Bad; stratified random sample, seed = 42) for all comparisons. The normative logistic regression on this subset achieves 75.1% accuracy and ROC AUC = 0.751, establishing the performance ceiling.

The org-externalized condition injects normative cue weights (derived from ridge logistic regression on the full 1,000-case dataset) as bank policy before each case, structured by tier (strong, moderate, weak) with direction. The introspective condition provides each model with a personalized deviation profile: which cues it over- and underweighted in the baseline condition relative to the normative model.

5.3. Baseline Process Alignment

Table 1 reports baseline results. Cosine similarity varies across models (range: -0.229 to $+0.503$), indicating heterogeneous implicit policies comparable in spread to the ECHR domain.

This variation does not, however, translate into accuracy differences. All five models achieve 44–54% accuracy on the balanced subset, leaving a substantial gap relative to the 75.1% normative ceiling. More tellingly, across all balanced model–condition pairs, the Pearson correlation between cosine similarity and output accuracy is $r = 0.15$ ($p = .60$), effectively zero. This stands in sharp contrast to the ECHR domain ($r = 0.85$, $p < .001$). In the credit domain, process alignment and output accuracy are orthogonal: knowing how closely a model's policy resembles the historical benchmark tells us nothing about its classification performance.

The Good% column reveals another anomaly: the implicit approval rates span 5.5% (DeepSeek) to 68% (GPT-5.4-mini), despite all models encountering the same balanced cases. DeepSeek and Claude Haiku classify almost all applicants as Bad risk; GPT-5.4-mini closely tracks the 70% historical base rate. These are fundamentally different implicit policies about credit risk, yet they produce nearly identical accuracy figures.

Table 1. Baseline process alignment in the German Credit domain (balanced subset, $n = 600$). Normative ceiling: 75.1% accuracy, AUC = 0.751. Good% = proportion of cases classified as Good risk.

Model	$r_{\cos}$	Acc (%)	AUC	Good%
Claude Haiku 4.5	$+0.503$	53.5	0.930	9.2
GPT-5.4-mini	$+0.060$	48.3	0.961	68.0
GPT-5.4-nano	$+0.499$	44.2	0.936	50.5
Grok 4.1 Fast	-0.229	48.8	0.882	37.5
DeepSeek-v3.2	$+0.264$	52.5	0.925	5.5

Table 2. Change in process alignment under org-externalization and introspective-externalization ($\Delta r_{\cos}$, German Credit balanced subset). †Grok introspective condition excluded: model classified 99.5% of cases as Good (degenerate over-correction).

	Org-ext		Introspective	
Model	$r_{\cos}$	Δ	$r_{\cos}$	Δ
Claude Haiku 4.5	-0.104	-0.607	$+0.360$	-0.143
GPT-5.4-mini	$+0.224$	$+0.164$	$+0.219$	$+0.159$
GPT-5.4-nano	$+0.076$	-0.423	-0.256	-0.755
Grok 4.1 Fast	$+0.176$	$+0.405$	—	†
DeepSeek-v3.2	$+0.164$	-0.100	-0.183	-0.447

5.4. Effect of Externalization

Table 2 reports alignment changes under both interventions. The pattern is inconsistent and, on average, negative.

Org-externalization moves two models toward the normative policy (GPT-5.4-mini $\Delta = +0.164$; Grok $\Delta = +0.405$) and moves three away (Claude Haiku $\Delta = -0.607$; GPT-5.4-nano $\Delta = -0.423$; DeepSeek $\Delta = -0.100$). Claude Haiku's sharp negative shift is particularly notable: providing the explicit normative policy as guidance *reduces* its alignment, suggesting that explicit credit scoring instructions push the model toward a different implicit reasoning pattern. Introspective externalization fares worse overall: three of four evaluable models decline in alignment (Δ range: -0.755 to $+0.159$).

Grok's degenerate over-correction. When given personalized deviation feedback noting its under-approval relative to the 70% historical base rate, Grok 4.1 Fast classified 99.5% of cases as Good risk in the introspective condition, effectively inverting its prior 37.5% approval rate. Rather than recalibrating toward the normative policy, Grok 4.1 Fast treated the base-rate instruction as a decision rule. This over-

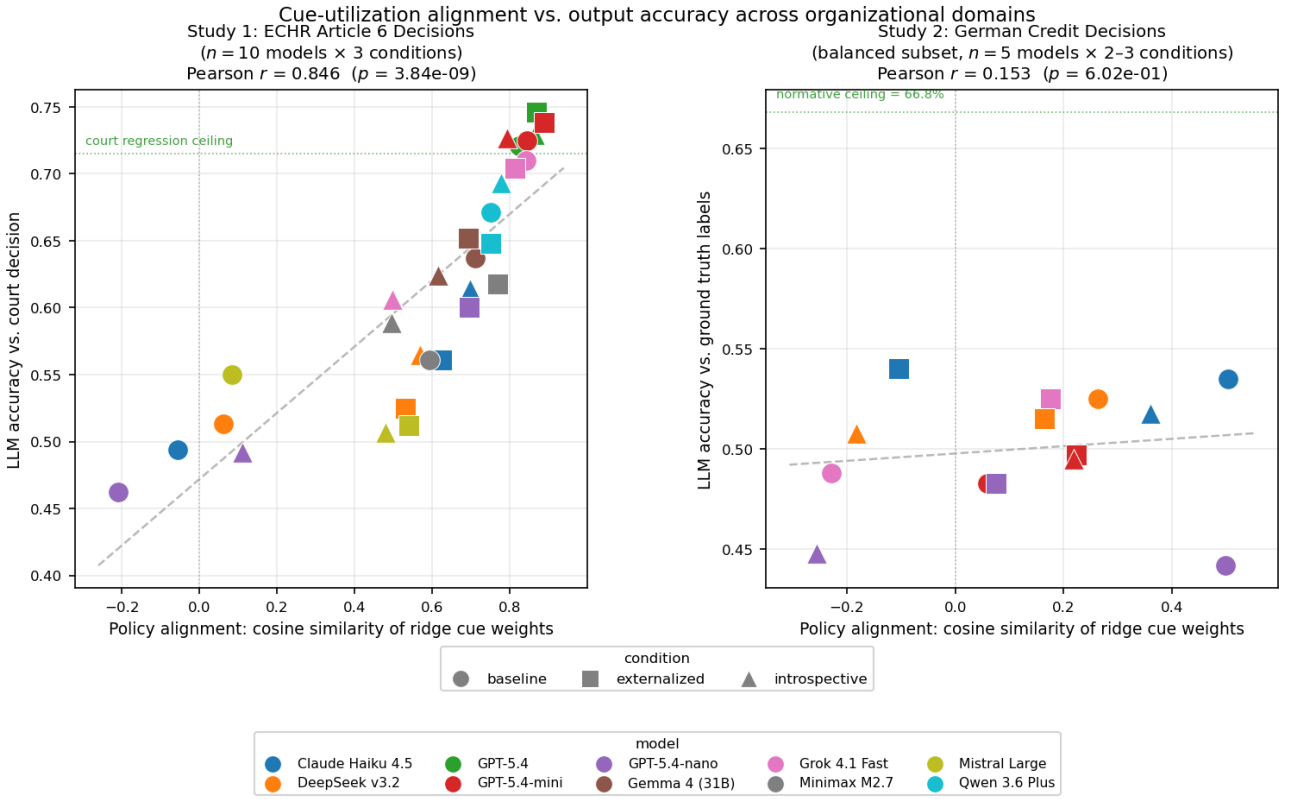

Figure 1. Policy alignment (cosine similarity) vs. output accuracy across both domains. **Left:** ECHR Article 6 decisions show a strong positive relationship ($r = 0.85$, $p < .001$, $n = 10$ models $\times$ 3 conditions); the dotted line marks the court regression ceiling. **Right:** German Credit decisions (balanced subset) show a flat, non-significant relationship ($r = 0.15$, $p = .60$, $n = 5$ models $\times$ 2–3 conditions); the dotted line marks the normative ceiling (75.1%). The contrast is the central finding: in a domain with a contested normative benchmark, process alignment and output accuracy decouple completely.

correction renders cosine similarity analysis uninformative for this cell and illustrates a failure mode of introspective calibration: models may respond to statistical feedback with behavioral compliance rather than genuine policy adjustment.

5.5. Protected Attribute Weighting

Inspection of the descriptive tier analysis reveals that LLMs systematically diverge from the normative model on protected attributes. In the baseline condition, Claude Haiku weights `foreign_worker` as HIGH in 56.8% of cases, despite the normative model rating it LOW, while classifying `age_years` and `personal_status_sex` as NOT_MENTIONED in 34.5% and 36.8% of cases respectively, against a normative MEDIUM weighting for both. Interventions partially shift these patterns but do not consistently align them with the normative benchmark. This points to a deeper issue: models appear to possess domain knowledge about protected attributes, including awareness that these should not serve as decision criteria, which competes with and may override the organizational policy signal. Process alignment is thus doubly contested in this domain: models may resist adopting the organizational policy precisely on the cues where it diverges most from contemporary fairness norms.

6. Discussion

6.1. The Contrast Is the Finding

Our two studies produce a sharp cross-domain contrast. In the ECHR domain, process alignment predicts output accuracy strongly ($r = 0.85$), models start from diverse baselines, and externalization reliably closes the gap for misaligned models. In the German Credit domain, the alignment-accuracy relationship collapses ($r = 0.15$), LLMs achieve uniformly poor accuracy despite varied alignment (44–54% against a 75.1% ceiling), and interventions produce inconsistent results.

We argue this contrast reflects a fundamental difference in the nature of the two organizational benchmarks. The ECHR's cue-weighting policy reflects accumulated legal doctrine: there is a relatively stable, publicly articulated stan-

dard that well-trained models can learn from pretraining data and that externalization can reinforce. German consumer credit decisions in the 1990s reflect a combination of legitimate risk signals (loan duration, credit amount, repayment burden) and historically contingent social judgments that have since been partially superseded by anti-discrimination law.

The result is that high process alignment with the German Credit normative model is not a clean target. A model that perfectly replicates the historical credit officer's cue-weighting policy may be accurately capturing credit risk signals, or faithfully reproducing discriminatory judgments about gender, age, and immigration status. CALM cannot adjudicate this; its contribution is to make the behavioral policy legible so that the question can be asked.

A natural objection is that the two domains were selected to produce this contrast, rendering the contestedness explanation post hoc. We address this directly. The domains were chosen on the basis of data availability, institutional prominence, and structural distinctiveness before results were examined. The contestedness hypothesis carries genuine empirical content: it predicts that the alignment-accuracy relationship will be strong where the normative benchmark reflects stable, publicly-articulated standards, and weak where it reflects historically contingent social judgments. This prediction would be falsified if a domain with clearly contested norms showed a strong alignment-accuracy correlation, or if a domain with a legitimate normative benchmark showed the orthogonality observed in German Credit. Testing CALM on additional domains—clinical guidelines, hiring rubrics, audit standards—would allow norm stability and discriminative ceiling to be varied independently, providing a more rigorous test of the mechanism.

6.2. Steerable Pluralism Requires Process-Level Measurement

Sorensen et al. (2024) argue that steerable pluralism is the most practically important mode of pluralistic alignment: systems should be steerable toward the perspective of a specified stakeholder when directed. Our externalization results demonstrate that this steerability is partial in the legal domain and largely absent in the credit domain. Residual divergence persists even with explicit policy guidance, consistent with the view that tacit organizational knowledge cannot be fully captured in a prompt (Polanyi, 1966).

Steerability is not, however, straightforwardly *desirable* in the credit domain. An organization might specify its historical decision policy only to find that it encodes attributes that models resist using precisely because they violate training-time alignment guidelines. The tension between organizational process alignment and model-level safety alignment is a distinctively pluralistic challenge that output-level evaluation cannot surface.

6.3. The Faithfulness Problem

CALM infers cue-weighting policy from observed decisions, treating the model as a black box. It is worth emphasizing that this approach does not depend on chain-of-thought reliability. The estimated β_{LLM} captures the actual cue-weighting function the model implements across the full case set, derived from input-output pairs alone. This is the behavioral ground truth, independent of what the model claims to be doing in any individual reasoning trace. If anything, the potential divergence between stated reasoning and behavioral policy strengthens the case for CALM: a model can produce coherent, on-policy justifications while behaviorally weighting entirely different cues, meaning that auditing stated reasoning is insufficient and behavioral measurement is essential.

A separate but related question is whether a model's explicit chain-of-thought reasoning, produced when justifying a decision, reflects the same policy as its behavioral decisions. Prior work suggests systematic divergence: models can produce plausible-sounding reasoning that does not correspond to the features that actually drove their output (Lanham et al., 2023; Turpin et al., 2024). In the organizational alignment context, this creates a faithfulness problem with practical stakes. A model that explicitly cites the right cues in its reasoning while behaviorally weighting different ones would pass a surface-level audit but fail a process-alignment audit. Conversely, a model might weight cues in a normatively appropriate way while producing reasoning that references protected attributes.

A natural extension of the present work is to compare CALM's behavioral policy estimates against cue weights extracted from explicit LLM reasoning, for instance by asking a scoring model to classify which cues were mentioned as influential in each case, then regressing those mention rates against the same cue set. If behavioral policy and stated reasoning track each other, then explicit reasoning is a reliable proxy for process alignment. If they diverge, auditing stated reasoning is insufficient and behavioral measurement via CALM is essential. We leave this analysis to future work.

6.4. Process Alignment as an Audit Tool

Our findings suggest a revised framing for CALM. In domains with legitimate, well-specified normative benchmarks (like ECHR), CALM is a *calibration* tool: it measures how far a model deviates from the target policy and guides externalization to close the gap. In domains with contested normative benchmarks (like German Credit), CALM is primarily an *audit* tool: it surfaces which protected and non-protected attributes each model is weighting, how those weights compare to the historical benchmark, and how they

shift under intervention. This diagnostic use does not require that the normative model be a legitimate alignment target; the comparison need only be informative.

For organizations deploying LLMs in high-stakes decision settings, CALM audits provide a third form of evaluation alongside output accuracy (are decisions correct?) and demographic disparity metrics (are decisions fair?): *are decisions reached in the right way?* Process legitimacy is increasingly recognized as a legal and ethical requirement in consequential AI deployments (Tyler, 1988), and CALM provides a scalable method for measuring it from observed decisions.

6.5. Policy Implications

Our findings have direct relevance to emerging AI governance frameworks. The EU AI Act classifies consumer credit scoring as a high-risk AI application (Annex III), requiring conformity assessments, transparency obligations, and human oversight. Current regulatory guidance focuses primarily on output-level metrics: accuracy, demographic parity, and calibration. Our results suggest this is insufficient. Two systems with identical accuracy and disparity profiles can differ dramatically in *how* they reach their decisions, with one replicating historically discriminatory weighting policies and the other not. CALM provides a process-level audit mechanism that can surface this divergence from observed decisions alone, without requiring access to model weights or internal representations.

More broadly, the question of *whose* organizational policy an LLM should align with is itself a governance question. Organizations deploying AI in regulated domains must specify not just a performance target but a normative target: which version of their decision policy should serve as the alignment reference, whether historical practice, stated ethical commitments, or current regulatory standards. CALM makes this choice explicit and auditable. We argue that pluralistic alignment frameworks should engage directly with these governance questions, since the technical problem of measuring alignment cannot be separated from the normative problem of deciding what to align to (Russell, 2019; Gabriel, 2020).

7. Conclusion

With CALM, we have applied a tool to measure process alignment across legal (ECHR) and financial (German Credit) decision domains. The results show that the relationship between process alignment and output accuracy is domain-dependent: strong and reliable where the organizational benchmark reflects stable, publicly-articulable norms; absent where it reflects historically contingent and contested values. In the credit domain, interventions that successfully shift model behavior in the legal domain fail to reliably improve alignment, and at least one model responds to statistical feedback with degenerate overcorrection. The credit domain also surfaces the fairness tension inherent in organizational alignment: models may actively resist adopting historical decision policies on protected attributes.

We conclude that organizational alignment is not a single problem. It requires asking: which organization, which policy, and from whose perspective? CALM provides a measurement framework for asking these questions empirically. We argue that pluralistic alignment research should extend beyond between-person value diversity to encompass between-organization diversity, the ineffability of tacit institutional knowledge, and the ethical contestability of the norms that organizational benchmarks encode.

Impact Statement

This work develops methods for measuring whether AI systems reason as specific organizations do. These methods have positive applications in auditing AI decision-making for process legitimacy and fairness. The same methods could, in principle, be used to align AI systems with organizations whose decision policies are discriminatory or harmful. We argue that surfacing this risk, as CALM does in the German Credit analysis, is preferable to leaving it invisible in output-level evaluations.

References

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., and Lampos, V. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93, 2016.

Binz, M. and Schulz, E. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6), 2023.

Brunswik, E. *The Conceptual Framework of Psychology*. University of Chicago Press, 1952.

Chalkidis, I., Fergadiotis, M., Tsarapatsanis, D., Aletras, N., Androutsopoulos, I., and Malakasiotis, P. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

Dolata, M., Feuerriegel, S., and Schwabe, G. A sociotechnical view of algorithmic fairness. *Information Systems Journal*, 32(4):754–818, 2022.

Gabriel, I. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437, 2020.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Hammond, K. R., Hursch, C. J., and Todd, F. J. Analyzing the components of clinical inference. *Psychological Review*, 71(6):438–456, 1964.

Hofmann, H. Statlog (German Credit Data). https://doi.org/10.24432/C5NC77, 1994. UCI Machine Learning Repository.

Karelaia, N. and Hogarth, R. M. Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, 134(3):404–426, 2008.

Lanham, M. et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Lebovitz, S., Levina, N., and Lifshitz-Assaf, H. Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *MIS Quarterly*, 45(3):1501–1526, 2021.

Nonaka, I. A dynamic theory of organizational knowledge creation. *Organization Science*, 5(1):14–37, 1994.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Pakarinen, P. and Huising, R. Relational expertise: What machines can't know. *Journal of Management Studies*, 62(5):2053–2082, 2023.

Polanyi, M. *The Tacit Dimension*. Doubleday, 1966.

Russell, S. Human compatible: Artificial intelligence and the problem of control. 2019.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 59–68, 2019.

Smith, L., Gilhooly, K., and Walker, A. Factors influencing prescribing decisions in the treatment of depression: A social judgement theory approach. *Psychology, Health & Medicine*, 17:51–63, 2003.

Sorensen, T., Bhatt, L., Gordon, M., Hashimoto, T., Moschella, L., Radford, A., Reimers, N., Smith, N. A., Talat, Z., and Zettlemoyer, L. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. 2024.

Tyler, T. R. *Why People Obey the Law*. Yale University Press, 1988.

Wieringa, M. What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 1–18, 2021.