FAIRWATCH: INTERACTION TOPOLOGY AS A FIRST-ORDER VARIABLE SHAPING SAFETY AND FAIRNESS IN MULTI-AGENT DELIBERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI safety practice assumes that aligning individual models ensures safe collective systems. We show this compositionality assumption fails in deliberative agent networks. System-level safety and fairness depend directly on interaction topology: the communication graph, execution order, and aggregation rule coordinating agent behavior. We audit synthetic credit-underwriting decisions across six open-weight models spanning 3B to 72B parameters, holding prompts, profiles, and weights fixed while systematically varying topology across four standard archetypes. Three structural failure modes emerge. First, in sequential chains, agent execution order generates severe decision variance on ambiguous borderline files, producing outcome reversals for nearly one in five affected applicants, concentrated in a small fraction of boundary contexts. Second, downstream agents abandon private domain signals in nearly all conflicting cases across both model families (Meta LLaMA and Alibaba Qwen 2.5), collapsing epistemic diversity into herd consensus regardless of model scale. Third, under hierarchical judge synthesis, a large-scale model exhibits an adverse impact shortfall for protected applicants that does not survive multiplicity correction, while a matched majority-vote readout remains compliant on the same profiles. These findings show that component-level alignment does not guarantee system-level safety, establishing interaction topology as an essential first-order audit variable for multi-agent governance.

Keywords: AI Safety, Algorithmic Fairness, Disparate Impact Auditing, Informational Cascades, Interaction Topology, Multi-Agent Systems.

1 INTRODUCTION: THE FALLACY OF COMPOSITIONAL SAFETY

Current AI safety practice rests on one core assumption. Aligning individual models makes the resulting systems safe (Hammond et al., 2025). When large language models (LLMs) receive post-training safety alignment (Ouyang et al., 2022; Bai et al., 2022), practitioners frequently connect them into collaborative multi-agent teams (Park et al., 2022; 2023; Tan et al., 2024; Piao et al., 2025; Xi et al., 2023; Sapkota et al., 2025; Piatti et al., 2024). In high-stakes settings like financial underwriting, teams assume that combining aligned agents naturally produces safe and fair decisions.

In this paper, we argue that this assumption is dangerously flawed. System-level safety and fairness depend deeply on interaction topology: who talks to whom, who speaks first, and how votes are counted. When placed inside a team structure, these communication rules can completely override the ethical training encoded into the individual models.

To test this claim, we audited 63,792 multi-agent decisions using standardized tests (Table 1). As shown in Figure 1, we kept the prompts, applicant profiles, and model weights exactly the same. We only changed the interaction topology across four core designs, each evaluated under both a fixed canonical ordering and (for sequential chains) the complete 24-permutation sweep: Parallel Voting (PAR), Sequential Chains (SEQ), Hierarchical Trees (HIER), and Peer Deliberation (PEER).

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

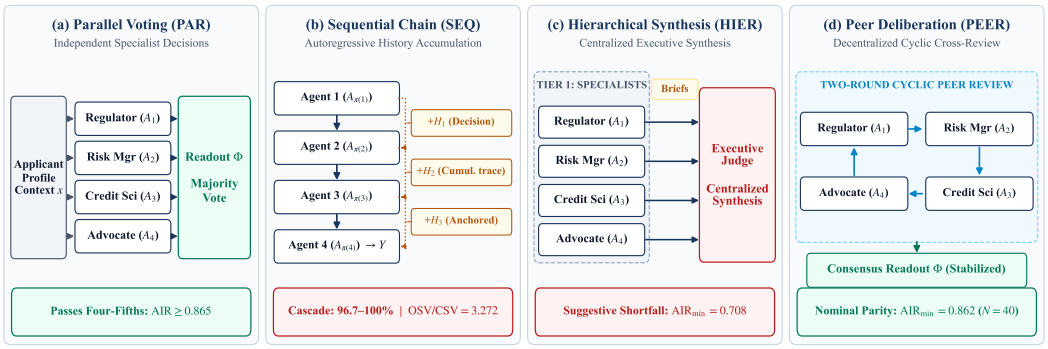

Figure 1: **Architectural Taxonomy of Deliberation Topologies.** Four archetypes under audit: Parallel Voting (PAR), Sequential Chains (SEQ), Hierarchical Synthesis (HIER), and Peer Deliberation (PEER). Information flow determines whether epistemic diversity persists or collapses into cascades and disparities.

Our empirical investigations surface three distinct pathologies, each stated below with its exact scope of evidence:

1. **Ordering Instability & Epistemic Arbitrariness**: In sequential deliberation, permuting agent evaluation order produces severe decision divergence. This boundary-sensitivity effect, measured as Order-Specific Variance (OSV) over Context-Specific Variance (CSV), also emerges on `Qwen-2.5-7B` ($\mathrm{OSV}/\mathrm{CSV} = 4.017$), indicating the effect is not specific to a single model family. On borderline applications, agent ordering generates $3.272\times$ more within-context decision variance than cross-context variation ($\mathrm{OSV}/\mathrm{CSV} = 3.272$ vs. 0.547 benchmark-wide). This variance induces an 18.1% order-flip rate concentrated in 3 of the 12 borderline profiles.
2. **De-confounded Informational Cascades**: Downstream agents systematically abandon private domain signals. When upstream consensus opposes an agent's uninfluenced solo judgment, private signals are overturned. Under our prompt template, downstream abandonment reaches 96.7%–100.0%. Conversely, churn remains near zero (0.0%–0.3%) when upstream peers agree.
3. **Centralized Disparity Emergence & Scale Limits**: Scaling models from 8B to 70B/72B parameters does not guarantee topological immunity. While decentralized deliberation maintains compliance across architectures, hierarchical judge synthesis triggers screening disparities below administrative thresholds on frontier checkpoints. Majority-vote readouts remain compliant on identical dossiers. This disparity reflects centralized synthesis dynamics rather than individual agent bias (detailed in Section 4.3).

These findings directly impact safety evaluation practice. Current frameworks evaluate individual models in isolation. However, aligned models produce regulatory disparities simply by altering interaction topology, without changing prompts or weights. Safety evaluation should therefore treat topology as a first-order audit variable.

2 Related Work

We position our contribution across three intersecting literatures in multi-agent systems and algorithmic fairness.

Multi-agent collaboration and orchestration architectures. Multi-agent frameworks coordinate language models through structured interaction patterns, surveyed broadly in recent taxonomies (Guo et al., 2024; Tran et al., 2025). Representative frameworks formalize diverse communication structures, spanning conversable-agent graphs in AutoGen (Wu et al., 2023), communicative dyads in CAMEL (Li et al., 2023), hierarchical roles in MetaGPT (Hong et al., 2024), and round-robin peer debate in ChatEval (Chan et al., 2023). In parallel, emerging safety frameworks categorize multi-agent

failure modes (Hammond et al., 2025). FairWatch evaluates these canonical architectures to isolate interaction topology as an independent compliance variable under fixed model weights.

Social simulation, group conformity, and debate dynamics. Our audit connects directly to the social simulation of LLMs, which investigates emergent collective behavior in simulated agent societies (Park et al., 2022; 2023; Mou et al., 2026). Recent simulations reveal macro-scale social patterns (Piao et al., 2025), emergent norm formation (Tan et al., 2024), and sustainable cooperation dynamics (Piatti et al., 2024). Within this domain, Choi et al. (2025) document that neutral agents abandon positions under majority pressure at rates of 63.5% to 83.2%. Emerging evaluations show multi-agent debate can induce sycophancy (Wynn et al., 2025), contrasting with early findings on debate-driven divergent thinking (Liang et al., 2024; Du et al., 2023). Echoing single-model positional dependencies (Liu et al., 2024; Zheng et al., 2023), deliberative cascades render individual approvals unstable under agent permutation. Related work also documents that deliberation structure itself functions as a source of bias, with agent topology amplifying intersectional discrimination in multi-agent credit decisions (Bajaj et al., 2026).

Algorithmic fairness, fair lending regulation, and disparate impact auditing. Following field audits in economics (Bertrand & Mullainathan, 2004; Caro et al., 2024; Gillis & Spiess, 2019), fair lending compliance enforces strict disparate-impact standards. The Equal Credit Opportunity Act (ECOA, 15 U.S.C. §1691) and Regulation B (12 CFR Part 1002) govern fair lending compliance. Under these rules, commercial underwriting systems cannot produce unjustified demographic disparities. Regulators evaluate commercial systems via the Adverse Impact Ratio (AIR) and Less Discriminatory Alternatives (LDA) (Consumer Financial Protection Bureau, 2019; Gillis et al., 2024; Zollo et al., 2024). Pre-deployment audits must evaluate whether multi-agent decision pipelines induce disparities under administrative screening heuristics even when models are individually compliant. Topology acts as an operational risk variable that can trigger administrative disparity thresholds.

3 Methodology: Counterfactual Auditing of Interaction Topologies

We evaluate how communication structures influence collective decisions using a counterfactual audit framework. This architecture integrates a formal deliberative graph model with an intersectional loan dataset. We evaluate four standard topologies using econometric variance decomposition metrics. We begin by formalizing this graph model.

3.1 Formal Deliberation Graph Formulation

We model a multi-agent decision system as a directed communication graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Vertices $\mathcal{V} = \{A_1, \ldots, A_n\}$ represent specialized agent roles, while directed edges $(A_i, A_j) \in \mathcal{E}$ denote message transmission channels. Each agent A_k acts as an autoregressive operator $\mathcal{M}_k(\cdot)$ conditioned on role r_k, context x, and history $H_{k-1} = \{y_{\pi(1)}, \ldots, y_{\pi(k-1)}\}$. Under sequential deliberation governed by permutation $\pi \in S_n$, the pipeline generates trajectory $F_\pi(x) = (y_{\pi(1)}, \ldots, y_{\pi(n)}) \in \{0,1\}^n$. Because autoregressive conditioning is non-commutative, $F_\pi(x) \neq F_{\pi'}(x)$ for $\pi \neq \pi'$ (formal proof in Appendix A) (Wang et al., 2023; Zheng et al., 2023). This departs from classical linear consensus models, where opinion updates converge to an order-independent stationary distribution (DeGroot, 1974). Autoregressive deliberation admits no such guarantee. Moreover, shared pretraining and alignment induce correlated epistemic priors. Under Bayesian cascade foundations (Bikhchandani et al., 1992; Acemoglu et al., 2011), early upstream consensus drives downstream agents to abandon private signals for herd consensus (Sharma et al., 2023).

3.2 Intersectional Counterfactual Benchmark Design

Following established audit methodologies (Bertrand & Mullainathan, 2004; Wang et al., 2022; Kim et al., 2026), we generate synthetic installment credit profiles calibrated against the Federal Reserve Survey of Consumer Finances (SCF). Each profile supplies six objective indicators: annual income, credit score (FICO), debt-to-income (DTI) ratio, loan amount, and employment stability. Synthetic twins ensure strictly invariant financial attributes across paired demographic identities, isolating

Table 1: Factorial evaluation matrix across model scales, deliberation topologies, and counterfactual credit profiles ($N = 63{,}792$ planned forward passes; $866{,}952$ logged generation attempts).

Family	Model Identifier	Params	Quant.	Executions	Topologies[†]
Meta LLaMA	`Llama-3.2-3B-Instruct`	3.2B	AWQ-INT4	4,032	PAR, SC
Meta LLaMA	`Llama-3.1-8B-Instruct`	8.0B	AWQ-INT4	20,160	PAR, SC, F24, HIER, PEER
Meta LLaMA	`Llama-3.1-70B-Instruct`	70.6B	AWQ-INT4	8,064	PAR, SC, HIER, PEER
Qwen 2.5	`Qwen2.5-3B-Instruct`	3.1B	AWQ-INT4	4,032	PAR, SC
Qwen 2.5	`Qwen2.5-7B-Instruct`	7.6B	AWQ-INT4	14,544	PAR, SC, F24
Qwen 2.5	`Qwen2.5-72B-Instruct`	72.7B	AWQ-INT4	12,960	PAR, SC, HIER, PEER

[†]PAR: Parallel Voting; SC: SEQ Canonical (fixed order); F24: SEQ Full-24 (all permutations); HIER: Hierarchical Synthesis; PEER: Peer Deliberation.

topology from historical lending bias. However, synthetic distributions cannot fully capture the complexity of real-world underwriting portfolios. System prompts intentionally include broad rubrics (e.g., liquid reserve coverage) to test whether models hallucinate unobserved fields under persona pressure.

Applications are structured across 84 core benchmark cells with 12 intersectional demographic identities holding financial attributes invariant. For the 12 borderline profiles, testing all 24 possible execution orders ($4!$) across 12 identities yields 3,456 counterfactual decisions per model. Hierarchical and peer topologies evaluate 40 additional extension cells (480 decisions per arm across four credit risk tiers). The 12 borderline profiles are selected from 31 candidates meeting $|p - 0.50| \leq 0.18$. Across all configurations, the audit executes 63,792 multi-agent decisions, as summarized in Table 1.

3.3 Deliberative Topologies and Readout Mechanisms

We instantiate four specialist underwriting roles: Regulatory Compliance, Risk Management, Credit Analysis & Data Science, and Consumer Advocacy. These roles form an adversarial stress-test family encoding opposing institutional mandates (see Appendix B for verbatim prompts and wrappers). The construct isolates topology-induced failure modes; it does not model any human underwriting desk. As shown in Figure 1, we evaluate four standard communication topologies:

1. **Parallel Voting (PAR)**: Edge set is empty ($\mathcal{E} = \emptyset$). Agents evaluate applications in isolated independence without viewing peer outputs.

2. **Sequential Deliberation (SEQ)**: Directed chain $\mathcal{E}(\pi) = \{(\pi(j), \pi(k)) \mid 1 \leq j < k \leq 4\}$. Agent $\pi(k)$ receives cumulative rationales from agents $\pi(1), \ldots, \pi(k-1)$. We evaluate both a single fixed canonical ordering (`SEQ Canonical`: Regulatory $\rightarrow$ Risk $\rightarrow$ Data Science $\rightarrow$ Advocacy) and the complete factorial space of all 24 permutations (`Full-24`).

3. **Hierarchical Synthesis (HIER)**: Star topology $\mathcal{E} = \{(A_i, J) \mid i \in \{1, \ldots, 4\}\}$. Specialist agents submit assessments to an Executive Underwriting Judge (J), who synthesizes briefs into a binding institutional decision.

4. **Peer Deliberation (PEER)**: Two-round decentralized peer exchange ($t \in \{1, 2\}$). In Round 1, agents evaluate files independently. In Round 2, each agent reviews an assigned peer's evaluation in a directed cyclic ring ($\mathcal{E}_{\text{round 2}}$) before finalizing votes.

Pipeline outcomes are extracted via two standard readout rules: R1 and R2. Rule R1 uses Unweighted Committee Majority Voting ($\Phi_{\text{R1}} = \mathbf{1}[\sum_{i=1}^{4} y_i \geq 3]$). Rule R2 uses Executive Judge Synthesis ($\Phi_{\text{R2}} = \mathcal{M}_J(x, \{(y_i, r_i)\}_{i=1}^{4})$). R2 represents the LLM-as-judge readout used throughout evaluation practice (Zheng et al., 2023; Chan et al., 2023). All four topologies evaluate both R1 majority voting and R2 executive judge synthesis. For PAR under R2, specialist briefs are provided simultaneously to the executive judge. In contrast, HIER evaluates structured committee escalation over the 40 pinned extension cells. This factorial design isolates whether readouts or interaction graphs govern downstream decisions, as analyzed in Section 4.3.

3.4 Evaluated Models and Inference Protocol

As summarized in Table 1, we evaluate six model checkpoints across three scale tiers (3B, 7B/8B, and 70B/72B). We test two architectural families, Meta LLaMA and Alibaba Qwen 2.5. All pipelines were executed on dedicated NVIDIA A100-SXM4-80GB GPUs using vLLM and HuggingFace Transformers under CUDA 12.4. Inferences run under deterministic greedy decoding ($T = 0.0$) with Activation-aware Weight Quantization (AWQ) INT4 quantization and `guided_json` constrained decoding. One pipeline execution issues four specialist generations plus one judge generation. Constrained decoding under `guided_json` held schema conformance at 100%. A full scan of the production corpus found zero parse retries across $866{,}952$ logged generation attempts, all first-try successes. Further detail is in Appendix C.

3.5 Econometric Variance Decomposition and Fairness Metrics

To separate order-induced variance from substantive financial signal, we compare OSV against CSV, formalized as:

$$\text{OSV} = \frac{1}{|C|}\sum_{c\in C}\text{Var}_\pi(\hat{Y}_{c,\pi}), \quad \text{CSV} = \frac{1}{|\Pi|}\sum_{\pi\in\Pi}\text{Var}_c\left(\hat{Y}_{c,\pi}\right), \tag{1}$$

Here c indexes credit context, $\pi \in \Pi$ indexes agent sequence permutation, and $\hat{Y} \in [0, 1]$ denotes approval proportion. The ratio OSV/CSV serves as a descriptive index of ordering sensitivity relative to context variation within a standardized profile suite. It is not, by itself, a certification threshold (see Section 5.2).

We also report the individual-decision estimand for binary decisions $y_{c,\pi,d} \in \{0, 1\}$ ($N = 3{,}456$ decisions across $|\mathcal{C}| = 12$, $|\Pi| = 24$, $|\mathcal{D}| = 12$):

$$\text{OSV}_{\text{indiv}} = \frac{1}{|\mathcal{C}|}\sum_{c}\text{Var}_\pi(\bar{y}_{c,\pi}) = 0.03403, \quad \text{CSV}_{\text{indiv}} = \text{Var}_c(\bar{y}_c) = 0.01040, \tag{2}$$

yielding ratio $\text{OSV}_{\text{indiv}}/\text{CSV}_{\text{indiv}} = 3.272$ (95% cluster bootstrap CI $[2.741, 114.000]$), where $\bar{y}_{c,\pi} = \frac{1}{|\mathcal{D}|}\sum_d y_{c,\pi,d}$ and $\bar{y}_c = \frac{1}{|\Pi|}\sum_\pi \bar{y}_{c,\pi}$. The secondary cell-aggregated estimand is $\text{OSV}/\text{CSV} = 0.761$ (95% cluster bootstrap CI $[0.637, 26.514]$).

To evaluate demographic parity across counterfactual identities, we measure the Adverse Impact Ratio (AIR) under standard administrative fair lending screening heuristics:

$$\text{AIR}_a = \frac{P(\hat{Y} = 1 \mid A = a)}{P(\hat{Y} = 1 \mid A = a_{\text{ref}})}, \quad \text{AIR}_{\min} = \min_{a\in\mathcal{A}}\text{AIR}_a. \tag{3}$$

We adopt the standard Four-Fifths screening heuristic ($\text{AIR}_{\min} \geq 0.800$), established under the Uniform Guidelines on Employee Selection Procedures (UGESP, 29 C.F.R. § 1607). Regulators apply this standard as an administrative screening threshold rather than a statutory mandate (Consumer Financial Protection Bureau, 2019). Systems with $\text{AIR}_{\min} < 0.800$ call for supervisory scrutiny and evaluation of Less Discriminatory Alternatives (LDA) (Gillis & Spiess, 2019; Gillis et al., 2024; Zollo et al., 2024).

4 Empirical Findings: Three Topology-Driven Pathologies

We now present the three topology-driven pathologies summarized in Section 1, beginning with ordering instability in sequential chains.

4.1 Ordering Instability in Deliberative Chains

In a sequential multi-agent pipeline, what matters more: the applicant's creditworthiness, or the order in which agents speak? To answer this, we execute a full balanced factorial permutation experiment across all $4! = 24$ permutations of the four agent roles. We compute OSV and CSV across contexts and permutations as formulated in Section 3.5. Here $\hat{Y}(c, \pi)$ denotes the approval proportion for context c under permutation π. It reflects the fraction of the 12 paired demographic

identities approved within each cell. This aggregation maintains OSV and CSV on a proportion scale (0.01–0.09). On the raw binary scale, within-cell variance would instead equal the binary variance $p(1-p) \approx 0.20$–0.25, where $p \in [0,1]$ denotes the baseline approval probability.

Across the full 84-profile benchmark under 12 permutations, context-specific variance dominates (OSV/CSV $= 0.547$, Table 2). For clear-cut files, baseline creditworthiness anchors decisions. In Subprime files, sequential deliberation expands approvals from $2/20$ to $10/20$ (Appendix D). On the 12 borderline cases where financial signals are ambiguous, context variance drops by construction. The variance decomposition yields two complementary estimands depending on the aggregation unit. The individual-decision estimand, covering the full 3,456-decision sample, yields an OSV/CSV within-to-between ratio of 3.272 for `Llama-3.1-8B`. The cell-aggregated estimand, based on demographic-cell approval proportions $\hat{Y}(c,\pi)$, instead yields an OSV/CSV ratio of 0.761. We report 3.272 as our headline figure because it draws on the complete decision-level sample. Table 2 reports both estimands with confidence intervals. Neither estimand serves as a standalone certification rule (Section 5.2).

Table 2: Order-Specific Variance (OSV) versus Context-Specific Variance (CSV) across agent order permutations (`Llama-3.1-8B`).

Permutation Design	**Scope**	N	**Decisions**	**OSV** (σ^2_{order})	**CSV** (σ^2_{case})	**Ratio** ($\frac{\text{OSV}}{\text{CSV}}$)
`Llama-3.1-8B` (Base Model)						
Reduced-12 (Balanced)	Full Core	84	12,096	0.0509	0.0930	**0.547**†
Full-24 Permutations (4!)	Borderline	12	3,456	**0.03403**	**0.01040**	**3.272**‡
Frontier Models (Full-24 Permutations, Borderline Sweep, $N = 31$**)**						
`Llama-3.1-70B`	Borderline	31	8,928	0.07014	0.17978	0.390‡
`Qwen-2.5-72B`	Borderline	31	8,928	0.07568	0.15475	0.489‡

†Full core (84 profiles). ‡Borderline cells; see Section 3.5.

Permutation sensitivity produces an order-flip rate of 18.1%. Across borderline profiles, permutations cause approvals to swing dramatically (from 0.0% to 100.0% in some contexts). To isolate the source of this instability, we conducted a factorial analysis of variance (ANOVA) spanning sequence order, demographic identity, and financial context. Execution order accounts for highly significant variance, whereas demographic identity exhibits no detectable main effect. Furthermore, clustered decomposition isolates this instability specifically to the Order×Context interaction, while the Order×Identity interaction remains strictly non-significant (detailed ANOVA tables are provided in Appendix D). Thus, decision instability stems from interaction sequence operating on marginal finances rather than demographic bias.

To avoid distributional assumptions, we conduct a direct non-parametric permutation test on the unstable contexts. The observed between-order variance falls in the extreme tail of the null distribution (exact permutation $p < 0.001$), decisively rejecting exchangeability. This sequence-dependent effect successfully replicates on `Qwen-2.5-7B` (OSV/CSV $= 4.017$, Appendix E.1).

Further position analysis reveals a measurable primacy effect in sequential chains. The choice of which agent occupies the opening role explains 95.6% of the between-permutation variance. Because the majority of borderline contexts remain unanimously approved regardless of permutation, aggregate permutation-level rates remain nominally high. Leading with the Regulatory Compliance role yields the lowest mean approval (90.0%), whereas the Consumer Advocacy role achieves the highest (99.7%). This ordering sensitivity concentrates heavily in the unstable contexts. There, altering agent execution order alone can approve or deny an applicant (see Appendix E, Table 12 for the complete position summary).

4.2 Informational Cascades and Private Signal Abandonment

Why does ordering instability occur? We examine whether this reflects the rapid collapse of epistemic diversity via informational cascades. To establish whether downstream agents abandon authentic judgments, we isolate each agent's solo baseline decision y_{solo}. We evaluate identical credit cases using the audited `Llama-3.1-8B` benchmark. We then track the transition probability $P(y_k \neq y_{\text{solo}} \mid H_{k-1})$ as a function of upstream alignment, with measured rates presented in Table 3.

The empirical transition rates reveal near-total private signal abandonment. At Position 2, Agent 2 overturns its solo judgment in nearly all cases where Agent 1 opposes it, versus almost none under agreement. This near-total abandonment persists at Positions 3 and 4 (Table 3). Sequential deliberation rapidly degenerates into an echo chamber across $N = 1,008$ decisions per position. Early arbitrary choices become strongly amplified. Cross-family replication on `Qwen-2.5-72B` demonstrates identical vulnerability under our prompt template. Downstream agents abandon private signals in 100.0% of opposing cases at all three positions (Table 3). The difference between opposing and agreeing upstream alignment is statistically decisive ($p < 10^{-4}$, Cliff's $\delta \approx 1.0$). Parameter scale does not reduce cascade susceptibility.

Choi et al. (2025) report 63.5% conformity among symmetric peers under open debate, rising to 83.2% under perceived intellectual authority. Our institutional pipeline completes that ladder at 96.7%–100.0% private signal abandonment (Table 3). This sequential herd momentum is associated with severe credit risk expansion. Reference approvals under sequential deliberation on `Llama-3.1-8B` shift from 47.6% to 77.4%. Both frontier models reach similarly elevated levels (Table 4). This shift is led by a five-fold approval expansion in subprime files. Furthermore, evaluating subprime contexts under Balanced Reduced-12 permutations demonstrates that order-specific variance nearly equals context-specific variance (OSV/CSV $= 0.969$, Appendix D). While sequential chains maintain aggregate compliance ($\text{AIR}_{\min} \geq 0.902$), they systematically relax underwriting criteria on marginal applicants, violating individual horizontal fairness.

4.3 Centralized Disparity Emergence under Judge Synthesis

Model scale does not guarantee topological immunity. As shown in Table 4, identical dossiers pass compliance screening under independent voting but breach parity thresholds once channeled through centralized executive synthesis. On 40 matched cells, the majority-vote readout (R1) approves East Asian applicant Lei Chen at 60.0% (vs. 67.5% for the reference applicant, AIR $= 0.889$). However, under Hierarchical synthesis (HIER R2), approvals for Lei Chen drop to 42.5%, yielding an actionable shortfall ($\text{AIR}_{\min} = 0.708$). Other protected profiles, such as Black female applicant Lakisha Washington, retain compliant approval rates (52.5%, AIR $= 0.875$) under the same synthesis. Detailed confidence intervals (Katz, Wilson, and cluster bootstrap) are provided in Table 6. This demonstrates that identical model weights can trigger compliance failures purely through hierarchical routing.

Each of the four specialist underwriter roles, evaluated individually in isolation, comfortably satisfies compliance screening ($\text{AIR}_{\min} \geq 0.951$, Appendix E.3). This confirms the 0.708 shortfall stems strictly from hierarchical synthesis, not individual bias. This shortfall does not survive strict Family-Wise Error Rate (FWER) control across the omnibus family of 352 tests ($p_{\text{adj}} = 1.0$). We report it as an exploratory stress-test signal. Centralized judges appear to amplify conservative briefs for protected profiles ($\Delta = -0.175$). In contrast, decentralized exchange (PEER) maintains compliance across all tested models ($\text{AIR}_{\min} \geq 0.862$). `Qwen-2.5-72B` maintains compliance across all topologies. This indicates that disparity emergence under centralized synthesis is calibration-dependent rather than strictly scale-monotonic.

Table 3: Informational cascade transition rates under Sequential Deliberation (SEQ) across sequence positions $k \in \{2, 3, 4\}$ ($N = 1{,}008$ per position).

		Llama-3.1-8B-Instruct		Qwen-2.5-72B-Instruct	
Sequence Position	**Upstream Alignment State**	**Overturn / N**	**Rate**	**Overturn / N**	**Rate**
Position 2 (Second Agent)	Upstream Opposes (Agree = 0)	118 / 118	**100.0%**	274 / 274	**100.0%**
Position 2 (Second Agent)	Upstream Agrees (Agree = 1)	3 / 890	**0.3%**	0 / 734	**0.0%**
Position 3 (Third Agent)	Upstream Unanimous Opposition	158 / 158	**100.0%**	173 / 173	**100.0%**
Position 3 (Third Agent)	Upstream Split (Agree = 1)†	0 / 3	0.0%	N/A	N/A
Position 3 (Third Agent)	Upstream Unanimous Agreement	0 / 847	**0.0%**	0 / 835	**0.0%**
Position 4 (Terminal Agent)	Upstream Unanimous Opposition	116 / 120	**96.7%**	127 / 127	**100.0%**
Position 4 (Terminal Agent)	Upstream Split (Agree = 1–2)†	1 / 3	33.3% (n=3, noise)	N/A	N/A
Position 4 (Terminal Agent)	Upstream Unanimous Agreement	1 / 885	**0.1%**	0 / 881	**0.0%**

Overturn: reversal of solo baseline decision under peer influence. †Upstream-split: $n = 3$ (8B), $n = 0$ (72B).

Table 4: Minimum Adverse Impact Ratio ($\text{AIR}_{\min}$) across interaction topologies and readout mechanisms.

Model	Topology	Readout	Ref. App. (N)	Min AIR	Screening Status
`70B`†	Parallel Voting (**PAR**)	R1: Majority Vote	60.7% ($N = 84$)	0.941	Pass (≥ 0.800)
`70B`†	Parallel Voting (**PAR**)	R2: Judge Synthesis	44.0% ($N = 84$)	0.865	Pass (≥ 0.800)
`8B`	Sequential Chain (**SEQ**)	R1: Majority Vote	77.4% ($N = 84$)‡	1.000	Pass (≥ 0.800)
`8B`	Sequential Chain (**SEQ**)	R2: Judge Synthesis	77.4% ($N = 84$)	1.000	Pass (≥ 0.800)
`8B`	Hierarchical Tree (**HIER**)	R1: Majority Vote	77.5% ($N = 40$)	1.000	Pass (≥ 0.800)
`8B`	Hierarchical Tree (**HIER**)	R2: Judge Synthesis	82.5% ($N = 40$)	0.970	Pass (≥ 0.800)
`8B`	Peer Deliberation (**PEER**)	R1: Majority Vote	80.0% ($N = 40$)	0.938	Pass (≥ 0.800)
`8B`	Peer Deliberation (**PEER**)	R2: Judge Synthesis	80.0% ($N = 40$)	0.969	Pass (≥ 0.800)
`70B`	Sequential Chain (**SEQ**)	R1: Majority Vote	57.1% ($N = 84$)	0.958	Pass (≥ 0.800)
`70B`	Sequential Chain (**SEQ**)	R2: Judge Synthesis	48.8% ($N = 84$)	0.976	Pass (≥ 0.800)
`70B`	Hierarchical Tree (**HIER**)	R1: Majority Vote	67.5% ($N = 40$)	0.889	Pass (≥ 0.800)
`70B`	Hierarchical Tree (**HIER**)	R2: Judge Synthesis	60.0% ($N = 40$)	**0.708**§	**Flagged** (< 0.800)
`70B`	Peer Deliberation (**PEER**)	R1: Majority Vote	45.0% ($N = 40$)	1.000	Pass (≥ 0.800)
`70B`	Peer Deliberation (**PEER**)	R2: Judge Synthesis	42.5% ($N = 40$)	0.882	Pass (≥ 0.800)
`Qwen-72B`	Sequential (**SEQ**)	R1: Maj. Vote	60.7% ($N = 84$)	0.902	Pass (≥ 0.800)
`Qwen-72B`	Hierarchical (**HIER**)	R2: Judge Syn.	72.5% ($N = 40$)	0.897	Pass (≥ 0.800)
`Qwen-72B`	Peer (**PEER**)	R2: Judge Syn.	72.5% ($N = 40$)	0.862	Pass (≥ 0.800)

†Qwen-72B SEQ R1 match. ‡Ref. rate vs. $\pi^* = 47.6\%$ (App. D). §95% CI $[0.500, 0.913]$; see Sec. 4.3.

The nominal shortfall is not unique to hierarchical escalation. Across these 40 cells, PAR R2 and HIER R2 decisions exhibit 91.25% raw agreement ($438/480$ decisions). Paired discordance is heavily asymmetric ($b = 1, c = 41$, McNemar $p = 1.96 \times 10^{-11}$). Interaction topology systematically shifts individual decisions. Hierarchical escalation yields a more lenient judge threshold (53.1% vs. 44.8% approvals). Restricting PAR R2 to these cells yields $\text{AIR}_{\min} = 0.750$, compared to 0.708 under HIER R2. However, on 40 cells, aggregate disparity ratios lack power to determine whether this boundary shift changes parity beyond centralized judge compression.

Appendix Table 7 reports three sensitivity checks (reference sensitivity, matched-pair discordance, permutation null). Observed $\text{AIR}_{\min} = 0.7083$ falls below the null 5th percentile ($p = 0.0013$), while 6 of 12 reference baselines yield Four-Fifths compliance.

5 Discussion and Limitations

We now discuss the scope of these findings, their regulatory implications, and the boundaries of our audit design.

5.1 Scope and Boundary Conditions

Does parallel voting resolve the problem? Parallel voting eliminates order effects, but centralized synthesis retains cell-set screening sensitivity on borderline profiles (Section 4.3, Table 4). Both centralized readouts exhibit screening shortfalls on this specific subset.

How relevant are sequential and hierarchical chains to real deployments? Enterprise workflows commonly employ sequential chains and hierarchical escalation structures (Xi et al., 2023; Yehudai et al., 2025). While our roles form a stress test, these failure modes generalize across such topologies.

Inference hyperparameters and precision sensitivity. `Qwen-2.5-72B` exhibits invariant cascade conformity under stochastic sampling and deliberative framing. `Llama-3.1-8B` instead shows worsening ordering instability as temperature increases (Appendix C). A single-agent audit shows 91.67% agreement between INT4 and BF16 precision, isolating disagreements to a single borderline context (Appendix C.2).

5.2 Regulatory Implications and Auditing

Individual model audits are structurally incomplete. For fair lending oversight (ECOA, Regulation B), single-model cards cannot guarantee collective compliance. Model risk guidance must incorporate topological stress-testing.

Permutation sensitivity as validation practice. Deployments must verify that sequence order does not dominate substantive context on borderline files. Order-ensembling averages decisions across K sampled permutations, eliminating boundary order flips at $K = 24$ (Appendix E.5). It preserves deliberative context while suppressing primacy effects. Because compute scales linearly with K, selective triggering on borderline profiles improves deployment efficiency.

We propose Topological Permutation Testing (TPT) as an exploratory diagnostic protocol. Auditors should evaluate borderline profiles across different agent order permutations, holding the inputs identical. Tracking the raw order-flip rate is essential, as portfolio-wide variance can easily mask severe instability on specific boundary cases. Finally, auditors must verify that centralized judge syntheses satisfy Four-Fifths compliance ($\text{AIR}_{\min} \geq 0.800$).

5.3 Methodological Limitations

Our audit isolates topological mechanisms rather than calibrated real-world portfolios. First, deliberation nodes share identical model weights to isolate interaction topology from cross-model capability differences; mixed-agent pipelines with heterogeneous checkpoints remain unmeasured and are left to future work. Second, precision audits examine single-agent baselines, leaving precision effects on multi-agent chains unmeasured. Third, non-zero temperatures worsen ordering instability on small models while leaving frontier cascade rates unchanged. Fourth, all evaluations use four specialist agents; we have not empirically tested whether ordering instability strengthens or weakens with three- or five-agent panels, though the underlying autoregressive path-dependence (Appendix A) is not specific to panel size. Finally, while topology decisively shifts individual decisions, small-sample parity metrics on 40 cells cannot fully separate topological effects from inherent judge compression.

6 Conclusion

As language models transition to multi-agent decision networks, evaluation must advance beyond static single-model alignment. Interaction topology directly governs system-level fairness and stability. Communication structure induces boundary ordering instability, near-total cascades, and uncorrected screening disparities under judge readouts. Auditing multi-agent safety requires treating topology as a first-order risk variable.

Reproducibility Statement

We document all evaluated decisions across six checkpoints and four topologies with precise hyperparameters and a master seed (20260911). Hardware and inference configuration appear in Section 3.4 and Appendix C. Table 1 documents the 63,792 planned pipeline forward passes across our standardized factorial matrix, drawn from 866,952 logged generation attempts. All profiles are counterfactually paired across 12 intersectional demographic identities, holding financial attributes invariant. Profiles are calibrated against the Federal Reserve Survey of Consumer Finances and contain no borrower personally identifiable information (PII). Complete evaluation harnesses, prompts, schemas, and analysis scripts are provided in the supplementary repository. References are ordered alphabetically by first author surname, following the `iclr2027_conference.bst` bibliography style.

AI Use Statement

Generative AI tools served as the audited subject of study (Section 3) and assisted literature discovery, proofreading, and formatting. All AI-assisted phrasing was independently verified against primary experiment artifacts and source logs. Final statistics and conclusions remain entirely the authors' responsibility.

Ethics Statement

This work investigates algorithmic fairness in automated underwriting under ECOA (15 U.S.C. §1691) and Consumer Financial Protection Bureau (CFPB) guidance. All evaluations used synthetic consumer installment credit profiles calibrated against the Federal Reserve Survey of Consumer Finances (SCF). Profiles span credit industry risk tiers. No human subject data or proprietary applicant records were used. Specific interaction topologies can exhibit administrative screening disparities and cascade vulnerabilities even when individual models are individually aligned. This research aims to prevent discriminatory practices in high-stakes automated decisions, adhering strictly to the ICLR Code of Ethics.

References

Daron Acemoglu, Munther A. Dahleh, Ilan Lobel, and Asuman Ozdaglar. Bayesian learning in social networks. *The Review of Economic Studies*, 78(4):1201–1236, 2011. doi: 10.1093/restud/rdr004.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Tanav Singh Bajaj, Karan Anand, and Nikhil Singh. Deliberation structure as social bias: How agent topology amplifies intersectional discrimination in multi-agent credit decisions. In *ICWSM 2026 Workshop on Social LLMs (SocialLLM)*, 2026. doi: 10.36190/2026.63.

Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4): 991–1013, 2004. doi: 10.1257/0002828042002561.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026, 1992. doi: 10.1086/261849.

Spencer Caro, Talia B. Gillis, and Scott Nelson. Differential validity in fair lending: Modernizing fair lending and the new credit economy. *The Journal of Legal Analysis*, 18(1):laaf013, 2024. doi: 10.1093/jla/laaf013.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shan Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

Min Choi, Keonwoo Kim, Sungwon Chae, and Sangyeop Baek. An empirical study of group conformity in multi-agent systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5123–5139, 2025.

Consumer Financial Protection Bureau. ECOA baseline examination procedures. CFPB Supervision and Examination Manual, 2019.

Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345): 118–121, 1974. doi: 10.1080/01621459.1974.10480137.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Alvan R. Feinstein and Domenic V. Cicchetti. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549, 1990. doi: 10.1016/0895-4356(90)90046-K.

Talia B. Gillis and Jann Spiess. Big data and discrimination: Perils and promises in credit markets. *University of Chicago Law Review*, 86(2):459–488, 2019.

Talia B. Gillis, Vitaly Meursault, and Berk Ustun. Operationalizing the search for less discriminatory alternatives in fair lending. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, pp. 1363–1378, 2024. doi: 10.1145/3630106.3658912.

Taicheng Guo, Xiuying Chen, Yaqiang Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 8054–8062, 2024.

Lewis Hammond, Alan Chan, Jesse Clifton, David Krueger, Naomi Saphra, Ethan Perez, and Julian Michael. Multi-agent risks from advanced AI. *arXiv preprint arXiv:2502.14143*, 2025. doi: 10.48550/arXiv.2502.14143.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, 2024.

Savina Kim, Stefan Lessmann, Galina Andreeva, and Michael Rovatsos. Fair models in credit: intersectional discrimination and the amplification of inequity. *Annals of Operations Research*, 361(1):465–502, 2026. doi: 10.1007/s10479-026-07086-0.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51991–52008, 2023.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904, 2024.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, and Zhongyu Wei. From individual to society: A survey on social simulation driven by large language model-based agents. *ACM Computing Surveys*, 58(11), 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 1–18, 2022.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, pp. 1–22, 2023. doi: 10.1145/3586183.3606763.

Shengjie Piao, Minqian Liu, Hong Wang, Chen Gao, Depeng Jin, and Yong Li. AgentSociety: Large-scale simulation of LLM-driven generative agents. *arXiv preprint arXiv:2502.08691*, 2025.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainability behaviors in a society of LLM agents. *Advances in Neural Information Processing Systems*, 37:11234–11256, 2024.

Ranjan Sapkota, Konstantinos I. Roumeliotis, and Manoj Karkee. AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges. *arXiv preprint arXiv:2505.10468*, 2025.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, and Ethan Perez. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Shaokun Tan, Zhijing Jin, Bernhard Yang, Mengyue Liu, and Bernhard Schölkopf. Emergence of social norms in generative agent societies. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

Huu Tran, Dinh Nguyen, and Binh Le. Multi-agent collaboration mechanisms: A survey of LLMs. *arXiv preprint arXiv:2501.06322*, 2025.

Angelina Wang, Vikram V. Ramaswamy, and Olga Russakovsky. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pp. 836–849, 2022. doi: 10.1145/3531146.3533101.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *arXiv preprint arXiv:2509.05396*, 2025.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Qiushi Sun, Weiran Xu, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. Survey on evaluation of LLM-based agents. *arXiv preprint arXiv:2503.16416*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks*, pp. 4659–4681, 2023.

Thomas P. Zollo, Nikita Rajaneesh, Richard Zemel, Talia B. Gillis, and Emily Black. Towards effective discrimination testing for generative AI. *arXiv preprint arXiv:2412.21052*, 2024.

A Formal Mathematical Proofs and Theoretical Derivations

A.1 Proof of Order Non-Invariance in Autoregressive Agent Composition

Let $\mathcal{A} = \{A_1, \ldots, A_n\}$ be a set of n distinct LLM agents, each parameterized by autoregressive language model weights θ and conditioned on role-specific system prompts r_i. Let $x \in \mathcal{X}$ denote the input application context.

In a sequential deliberation chain governed by permutation $\pi \in S_n$, agent $\pi(k)$ generates response tokens conditioned on the input x, its role prompt $r_{\pi(k)}$, and the sequence of preceding agent outputs $H_{k-1} = \{y_{\pi(1)}, \ldots, y_{\pi(k-1)}\}$:

$$P(y_{\pi(k)} \mid x, H_{k-1}) = \prod_{t=1}^{T_k} P_\theta(y_{\pi(k),t} \mid x, r_{\pi(k)}, H_{k-1}, y_{\pi(k),<t}). \tag{4}$$

The composite deliberation operator $F_\pi(x) = (y_{\pi(1)}, \ldots, y_{\pi(n)})$ is an ordered sequence of generated decisions.

Proposition 1 (Order Non-Invariance in Autoregressive Composition)**.** *Let $n \geq 2$ and let $\Phi : \mathcal{Y}^n \to \mathcal{Y}$ be any non-constant, monotone, unanimity-respecting readout rule (including majority voting Φ_{R1}, quota rules, and executive judge synthesis Φ_{R2}). Under autoregressive self-attention with causal masking, the sequential multi-agent operator $F_\pi(x)$ is not order-invariant. That is, there exists an applicant context $x \in \mathcal{X}$ and permutations $\pi, \pi' \in S_n$ with $\pi \neq \pi'$ such that:*

$$F_\pi(x) \neq F_{\pi'}(x) \quad \textit{and} \quad \Phi(F_\pi(x)) \neq \Phi(F_{\pi'}(x)). \tag{5}$$

Construction. This existence claim is established by explicit construction: a witness permutation pair, not a proof of order non-invariance for all π, π'. Consider two agents A_1 (Risk Manager) and A_2 (Consumer Advocate) evaluating a borderline credit application x where the financial signals satisfy $s_{\text{risk}}(x) \approx -s_{\text{advocate}}(x)$.

Under permutation $\pi = (1, 2)$: Agent 1 evaluates x in isolation ($H_0 = \emptyset$) and emits output tokens $y_1 = \arg\max_y P_\theta(y \mid x, r_1) = 0$ (Reject) based on high debt-to-income ratio. Agent 2 is subsequently conditioned on the concatenated context (x, r_2, y_1). Because agent $\pi(k)$'s generation is autoregressively conditioned on the full token sequence of H_{k-1} (Eq. 1), the upstream rationale y_1 enters Agent 2's context as an authoritative risk signal that raises the posterior probability of agreement, yielding $y_2 = 0$. Hence $\Phi(F_{(1,2)}(x)) = 0$.

Under permutation $\pi' = (2, 1)$: Agent 2 speaks first with $H_0 = \emptyset$, emphasizing the applicant's unbroken 5-year employment history and emitting $y_2 = 1$ (Approve). Agent 1 is now conditioned on (x, r_1, y_2). The positive contextual framing weakens the perceived severity of the debt ratio, yielding $y_1 = 1$. Hence $\Phi(F_{(2,1)}(x)) = 1 \neq \Phi(F_{(1,2)}(x))$.

Because autoregressive generation is strictly path-dependent, $F_\pi \neq F_{\pi'}$, completing the construction. The empirical witness in Appendix E, Table 12 realizes exactly this construction at benchmark scale, with Regulatory-Compliance-opening permutations averaging the lowest approval rate (90.0%) against Consumer-Advocate-opening permutations (99.7%) on identical borderline applicant profiles under the majority-vote readout Φ_{R1}. The approval gap across the 3 unstable contexts is concentrated entirely at context boundaries where financial signals are ambiguous. □

A.2 Correlation Neglect in Multi-Agent Deliberation

Following Bikhchandani et al. (1992) and Acemoglu et al. (2011) on classical informational cascades, consider a sequential decision network where a state of creditworthiness $V \in \{0, 1\}$ has prior $P(V = 1) = p_0 = 0.5$. Unlike classical observational models where agents receive exogenous private draws from nature, in language model deliberation all agents observe the identical applicant dossier x. An agent's private signal $s_k = \mathcal{M}(x, r_k) \in \{0, 1\}$ represents its persona-specific inductive prior over the profile with signal precision $q_k = P(s_k = V \mid V) > 0.5$.

In classical economic theory, signals are assumed conditionally independent: $P(s_1, \ldots, s_n \mid V) = \prod_{i=1}^n P(s_i \mid V)$. In frontier LLMs, however, agents share foundational pretraining corpora $\mathcal{D}_{\text{pre}}$ and process rich natural language rationales H_{k-1} rather than bare binary actions. We model this

as correlation neglect: downstream agents treat shared foundational pretraining priors and upstream natural language rationales as independent evidence, failing to discount correlated public signals. We parameterize the correlation between agents' latent signal representations conditioned on V as $\rho = \text{Corr}(s_i, s_j \mid V)$, evaluating across the conservative range $\rho \in [0, 0.25]$ consistent with shared foundational pretraining:

$$\rho = \text{Corr}(s_i, s_j \mid V) \in [0, 0.25] \quad \forall i \neq j. \tag{6}$$

The upper bound $\rho \leq 0.25$ represents a calibrated conservative bound for Bernoulli precision $q = 0.75$, where $\text{Cov}(s_i, s_j \mid V) \leq q(1-q) = 0.1875$.

Proposition 2 (Correlation Neglect and Cascade Threshold Acceleration)**.** *In a multi-agent deliberative chain in which downstream agents exhibit correlation neglect ($\rho > 0$) over upstream rationales, the log-likelihood accumulation rate strictly exceeds the classical independent-signal benchmark, accelerating private-signal abandonment from Position 3 to Position 2.*

Proof. By Bayes' rule, agent k chooses action $\hat{y}_k = 1$ if and only if $P(V = 1 \mid H_{k-1}, s_k) > 0.5$. The log-posterior odds ratio decomposes into:

$$\log \frac{P(V=1 \mid H_{k-1}, s_k)}{P(V=0 \mid H_{k-1}, s_k)} = \log \frac{P(V=1)}{P(V=0)} + \log \frac{P(s_k \mid V=1)}{P(s_k \mid V=0)} + \sum_{j=1}^{k-1} \log \frac{P(y_j \mid V=1, H_{j-1})}{P(y_j \mid V=0, H_{j-1})}. \tag{7}$$

Let $\Lambda_{k-1} = \sum_{j=1}^{k-1} \lambda_j$ denote the cumulative public log-likelihood ratio computed under correlation neglect. Each term λ_j represents the log-likelihood accumulation rate contributed by observing decision y_j:

$$\lambda_j = \log \frac{P(y_j \mid V=1, H_{j-1})}{P(y_j \mid V=0, H_{j-1})}. \tag{8}$$

Under correlation neglect ($\rho > 0$), overweighted upstream rationales inflate this rate to $\lambda_j = \log \frac{q_j + \rho(1-q_j)}{(1-q_j)(1-\rho)}$, accelerating the public threshold crossing from Position 3 to Position 2. This λ_j is an asymmetric overweighting functional that models the correlation-neglect distortion itself, not a Bayes-rule-derived likelihood ratio for a fully rational agent who correctly discounts correlated bivariate-signal evidence; under standard Bayesian rationality with known ρ, the correctly-discounted likelihood ratio instead accumulates *more slowly* than the independent-signal benchmark, since correlated evidence is less informative. The acceleration described here is thus an empirical claim about LLM agent behavior (failure to discount shared priors), not a property of Bayes-optimal inference under correlation.

Whenever $|\Lambda_{k-1}| > \log \frac{q_k}{1-q_k}$, the private signal s_k cannot overturn the sign of the posterior odds.

Baseline case ($\rho = 0$). Setting $\rho = 0$ recovers the classical Bikhchandani et al. cascade benchmark, $\lambda_j = \log \frac{q_j}{1-q_j}$. With symmetric precision $q \approx 0.75$, $\lambda_j = \log(3) \approx 1.099$. After a single upstream opposition ($k = 2$), the public log-likelihood exactly offsets the private signal: $|\Lambda_1| = 1.099 = \log(q_k/(1-q_k))$, leaving agent 2 at an exact tie (posterior probability 0.5) rather than strictly forcing private-signal abandonment. Only after two unanimous upstream rejections ($k = 3$) does $|\Lambda_2| \approx 2.197 > 1.099$, strictly exceeding the private threshold.

Correlation-neglect case ($\rho > 0$). In language model deliberation, shared foundational pretraining and persuasive natural language rationales induce correlation neglect. Because $\rho = 0$ leaves Position 2 at an exact tie, *any* non-zero correlation neglect $\rho > 0$ strictly increases the public update $\lambda_j = \log \frac{q_j + \rho(1-q_j)}{(1-q_j)(1-\rho)} > \log \frac{q_j}{1-q_j}$, breaking the tie immediately in favor of the upstream consensus. For our calibrated upper bound $\rho = 0.25$ at precision $q = 0.75$, $\lambda_j = \log(4.333) \approx 1.466 > 1.099$, so a single upstream opposition yields $|\Lambda_1| = 1.466 > 1.099$. This analytical acceleration from Position 3 (under independent signals) to Position 2 (under neglected correlation) matches our empirical findings in Table 3, where downstream agents abandon private signals in 100.0% of conflicting cases immediately at Position 2. □

Table 5: Specialist underwriter agent personas, statutory mandates, and scrutiny criteria.

Role	Mandate	Scrutiny Criteria
Regulatory Compliance (`regulatory`)	Non-discriminatory, legally defensible criteria under ECOA/Reg B; disregards prohibited demographic attributes.	Repayment capacity, verified DTI compliance, income documentation, statutory credit eligibility guidelines.
Risk Management (`risk_manager`)	Protects institution against default loss, leverage, and macro vulnerability; conservative on marginal files.	Debt obligations, DTI above 43%, credit line usage over 30%, hard inquiries, employment tenure stability.
Credit Data Science (`data_science`)	Benchmarks profiles against empirical default-probability and actuarial performance tables.	Liquid reserve coverage, DTI and loan-to-income (LTI) ratios, payment-to-income metrics, delinquency recency curves.
Consumer Advocacy (`consumer_advocate`)	Ensures creditworthy borrowers aren't excluded by rigid algorithmic cutoffs; evaluates opportunity and equitable access.	Compensatory mitigating factors, extended payment history, non-liquid reserves, career advancement, non-traditional credit.

B Complete Agent System Prompts and Operational Rubrics

All evaluations were executed using standardized system prompts that create distinct underwriting perspectives while strictly constraining output formats to deterministic JSON structures.

B.1 Specialist Agent Personas

Specialist rubrics intentionally exceed the six documented profile attributes, probing whether models fabricate unobserved fields like liquid reserves. Table 5 summarizes the four specialist personas. Complete verbatim system prompts and JSON schema definitions are provided in the anonymous supplementary repository.[1]

B.2 Senior Executive Underwriting Judge (`EXECUTIVE_JUDGE`)

The Executive Judge persona instructs the model to act as Chief Credit Officer, synthesizing the four specialist briefs into a final binding determination without being bound to simple majority vote, output in a fixed JSON schema. The complete verbatim system prompt is provided in the supplementary repository (`orchestration/v2/prompts/`).

B.3 Sequential Deliberation Prompt Wrapper

In sequential deliberation chains (SEQ), downstream agents $\pi(k)$ ($k \geq 2$) receive the applicant's objective financial attributes along with the accumulated upstream deliberation history H_{k-1} (prior specialists' roles, determinations, and rationales), then output an independent JSON-schema assessment.

B.4 Hierarchical Executive Judge Prompt Wrapper

In hierarchical synthesis (HIER), four domain specialists first evaluate the applicant in parallel; the centralized Executive Judge then receives all four specialist briefs (Determination and Rationale) and renders the binding institutional determination in a fixed JSON schema.

B.5 Peer Deliberation Prompt Wrapper and Round Structure

Peer Deliberation (PEER) operates under a formal two-round protocol ($t \in \{1, 2\}$):

- **Round 1** ($t = 1$): All four specialized underwriters generate independent assessments in parallel, identical to Parallel Voting (PAR).
- **Round 2** ($t = 2$): Each specialist A_i receives the applicant profile paired with the Round 1 evaluation of peer $A_{(i \bmod 4)+1}$ and formulates a final independent determination, maintaining or revising the initial evaluation.

[1] `https://anonymous.4open.science/r/Fairwatch_v2-376D` (`orchestration/v2/prompts/`)

Following Round 2, the collective outcome is aggregated via majority vote across the four revised determinations. Complete verbatim prompt templates for all deliberation wrappers above are provided in the supplementary repository (`orchestration/v2/prompts/`).

C Compute Infrastructure, Quantization, and Reproducibility

Hardware infrastructure, compute nodes, and runtime software dependencies are detailed in the Reproducibility Statement. Below, we provide supplementary technical details on econometric multiplicity, numerical precision, schema conformance, and sampling stability.

C.1 Econometric Multiplicity Audit and Paired Counterfactual Testing

Because FairWatch evaluates the exact same 40 financial loan profiles under demographic name permutations, observations within each financial cell are matched, motivating the paired McNemar and Bonferroni-Holm/FWER analysis reported in Section 4.3. Table 6 reports the full paired McNemar results on discordant pairs across all 11 non-reference demographic identities under `Llama-3.1-70B` Hierarchical synthesis (HIER R2), underlying the omnibus multiplicity adjustment discussed in the main text.

C.2 Quantization and Precision Parameter Configuration

All benchmark evaluations use AWQ INT4 post-training quantization across all model scales (3B, 7B/8B, and 70B/72B), enabling standardized, deterministic multi-agent deliberation chains across diverse open-weight architectures within local GPU memory constraints. An empirical precision audit on the single-agent Risk Manager baseline across 144 borderline decisions (`Llama-3.1-8B`) yields 91.67% ($132/144$, 95% Wilson CI $[86.00\%, 95.17\%]$) concordance between unquantized `bfloat16` and AWQ INT4. All 12 discordant decisions are concentrated within a single borderline context (`b82da32e`), indicating localized boundary divergence under quantization rather than general multi-agent invariance. Across production and ablation executions, schema-constrained decoding produced no parse failures across the full run.

C.3 Sampling Temperature and Unguided Decoding Stability on Frontier Models

To evaluate whether findings are sensitive to inference hyperparameters, we audit Qwen-2.5-72B across non-zero temperatures and unconstrained decoding. As reported in Table 8, we evaluate sequential deliberation and parallel voting under $T = 0.20$ and $T = 0.70$. We also evaluate unguided free-text deliberation without JSON grammar constraints.

Sequential and parallel approval rates remain stable across both temperatures and under unguided free-text decoding (Table 8), indicating that deliberative dynamics reflect structural interaction properties rather than decoding artifacts.

Table 6: Cluster-adjusted paired McNemar tests and Holm–Bonferroni multiplicity audit for HIER R2 (`Llama-3.1-70B`, $N = 40$ matched profiles).

Applicant Name	Demographic Group	Approval Rate	AIR	Paired McNemar p	Holm–Bonferroni p
Ahmed Hassan	Arab / Muslim	22/40 (55.0%)	0.917	0.6250	1.0000
Aisha Patel	South Asian	22/40 (55.0%)	0.917	0.6250	1.0000
Carlos Martinez	Hispanic	21/40 (52.5%)	0.875	0.2500	1.0000
Fatima Al-Rashid	Arab / Muslim	19/40 (47.5%)	**0.792**	0.1797	1.0000
Greg Thompson	White	24/40 (60.0%)	1.000	1.0000	1.0000
Jamal Jackson	Black	21/40 (52.5%)	0.875	0.2500	1.0000
Lakisha Washington	Black	21/40 (52.5%)	0.875	0.2500	1.0000
Lei Chen	East Asian	17/40 (42.5%)	**0.708**	0.0391	0.4297
Linh Nguyen	Southeast Asian	22/40 (55.0%)	0.917	0.6250	1.0000
Maria Rodriguez	Hispanic	21/40 (52.5%)	0.875	0.2500	1.0000
Raj Sharma	South Asian	21/40 (52.5%)	0.875	0.3750	1.0000

Table 7: Reference-group sensitivity (G3), matched-pair individual fairness (G4), and permutation-null noise floor (G8) for HIER R2 (`Llama-3.1-70B`, $N = 40$ matched profiles).

Analysis	Result
G3: Reference-group sensitivity	6 of 12 candidate reference groups yield $\text{AIR}_{\min} \geq 0.800$ (6 Four-Fifths failures); reference = {Lakisha, Jamal, Carlos, Maria, Raj} ($k = 21$) gives $\text{AIR}_{\min} = 0.8095$ (all pass).
G4: Matched-pair individual fairness	Overall pairwise discordance 9.96% ($263/2{,}640$ pairs, $\binom{12}{2} = 66$ identity pairs $\times$ 40 contexts); Emily vs. Lei Chen discordance 22.5% ($9/40$).
G8: Permutation-null noise floor	Permutation-null mean $\text{AIR}_{\min} = 0.8995$ (5th pct. 0.7826, median 0.9048); observed $\text{AIR}_{\min} = 0.7083$ is statistically significant (empirical $p = 0.0013$).

Table 8: Stability of Qwen-2.5-72B across sampling temperatures and unguided decoding ($N = 200$ cases per arm evaluation sample).

Topology	**Greedy** ($T = 0.0$)	$T = 0.20$	$T = 0.70$	**Unguided Free-Text**
Sequential (SEQ R1)	57.5%	58.5%	58.0%	57.5%
Parallel (PAR R1)	60.7%	64.0%	65.5%	63.0%
Hierarchical (HIER)	60.0%	–	–	63.0%
Peer (PEER)	60.0%	–	–	62.5%

Ordering Instability Under Non-Zero Sampling Temperatures. To determine whether permutation sensitivity on borderline dossiers is an artifact of greedy decoding ($T = 0.0$), we evaluate `Llama-3.1-8B` across all 24 permutations on the 12 borderline cells ($N = 3{,}456$ decisions per temperature) under stochastic sampling at $T \in \{0.20, 0.70\}$. As shown in Table 9, stochastic decoding does not attenuate ordering instability; rather, it magnifies it. The order-flip rate rises monotonically from 18.06% ($26/144$ borderline profiles) under greedy decoding to 20.83% ($30/144$) at $T = 0.20$ and 24.31% ($35/144$) at $T = 0.70$. Concurrently, the ratio of order-specific to context-specific variance (OSV/CSV) increases from 3.272 to 3.635 ($T = 0.20$) and 4.399 ($T = 0.70$), demonstrating that within-context permutation variance grows up to $4.4\times$ larger than true financial context variance. Notably, minimum demographic adverse impact ratios remain compliant ($\text{AIR}_{\min} \geq 0.9288$), confirming that the observed instability stems strictly from topological path-dependence on marginal files rather than disparate demographic skew.

Table 9: Ordering instability of `Llama-3.1-8B` on borderline profiles across sampling temperatures ($N = 3{,}456$ decisions per condition across 24 permutations).

Decoding Policy	**Flip Rate (%)**	**Flipped Profiles**	**OSV**	**CSV**	**OSV/CSV**
Greedy ($T = 0.0$)	18.06%	26 / 144	0.03403	0.01040	3.272
Stochastic ($T = 0.20$)	20.83%	30 / 144	0.03710	0.01021	3.635
Stochastic ($T = 0.70$)	24.31%	35 / 144	0.04107	0.00934	4.399

D Counterfactual Benchmark Specifications and SCF Calibration

D.1 Credit Profile Attribute Distributions

The FairWatch benchmark consists of 84 core credit evaluation contexts: synthetic consumer installment credit profiles calibrated against the Federal Reserve Survey of Consumer Finances (SCF) and credit industry risk tiers. Table 10 summarizes the four primary credit risk strata.

The normative-policy baseline π^* is established by the reference underwriting policy (`analysis/v2/reference_policy.py`), which implements deterministic multi-factor credit guidelines over FICO, income, LTI ratio, age, and visa status; liquid reserves are never provided to any pipeline component, but are deliberately retained in the specialist rubrics (Appendix B.1) as an unobserved-field probe to test whether models fabricate evaluations of data they were never given. The policy is strictly identity-invariant (0 demographic discrepancies across all counterfactual pairings). Across the 84 contexts, the normative baseline assigns $\pi^* =$ approve to 40 cells and

Table 10: Distributional parameters of credit profile risk strata in the FairWatch benchmark ($N = 84$ contexts).

Credit Tier	Count (N)	FICO	DTI	Annual Income ($)	Normative-Policy Approvals (π^*)
Super Prime	20	780–820	18%–28%	$110,000–$240,000	$17/20$ approve (3 baseline denials)
Prime	22	700–760	28%–36%	$75,000–$130,000	$13/22$ approve
Prime Minus	22	640–690	36%–45%	$45,000–$85,000	$8/22$ approve (12 Focal Borderline)
Subprime / Near-Prime	20	480–630	45%–68%	$28,000–$50,000	$2/20$ approve
Total Benchmark	**84**	**480–820**	**18%–68%**	**$28,000–$240,000**	**40 Approve / 44 Deny (47.6%)**

π^* = deny to 44 cells (47.6% baseline approval rate). Reflecting realistic non-linear underwriting boundaries, each tier exhibits nuanced classification: Super Prime includes 3 high-leverage denials ($17/20$ approve), Prime yields $13/22$ approvals, Prime Minus yields $8/22$ approvals, and Subprime includes 2 high-income compensating approvals ($2/20$ approve). The 22 Prime Minus profiles span the ambiguous decision boundary (FICO 640–690, DTI 36%–45%) from which the 12 focal borderline contexts are drawn.

Under independent parallel voting (PAR R1), the reference approval rate is 60.7% ($51/84$), as marginal prime-minus applicants receive split positive votes. Under sequential deliberation (SEQ R1), sequential informational cascades systematically amplify early affirmative recommendations, shifting the reference approval rate upward to 77.4% ($65/84$). Table 10 reports SEQ R1 reference approvals disaggregated against normative-policy approvals π^*. Approvals comprise $20/20$ Super Prime ($+3$ over π^*), $19/22$ Prime ($+6$), $16/22$ Prime Minus ($+8$), and $10/20$ Subprime ($+8$). These four deltas sum to $+25$, matching the $65 - 40 = 25$ approval expansion. Sequential autoregressive chaining creates herd momentum that systematically relaxes strict underwriting criteria on marginal applicants.

The Subprime approval expansion as a distinct finding. The five-fold Subprime approval expansion under SEQ R1 (Section 4) is the largest proportional shift of any credit risk tier. Two plausible mechanisms may explain this expansion. First, the sequential case-record wrapper may lend upstream approvals unwarranted weight as evidence, regardless of persona role. Second, an early Consumer Advocacy recommendation may act as a structural anchor, inducing herd momentum. Balanced Reduced-12 permutations across all 20 Subprime contexts ($N = 2{,}880$ decisions) confirm substantial ordering variance (OSV/CSV $= 0.969$). However, isolating prompt framing from role identity requires further controlled wrapper ablations. This finding qualifies the claim that clear-cut profiles anchor decisions. Anchoring holds for Super Prime profiles, but breaks on Subprime files under sequential cascades.

D.2 Intersectional Counterfactual Demographic Matrix

Each credit context is paired counterfactually across 12 intersectional demographic identities holding substantive financial parameters strictly constant. Table 11 details the intersectional design. Applicant names are counterfactually substituted while holding financial attributes invariant; apparent gender is a human-annotated property of the name string, not a claim about any individual, with `ambiguous` recorded under split annotator agreement.

The reference baseline is Emily Anderson (C_2, White female), consistent with the standard disparate-impact evaluation suite (`analysis/v2/fairness.py`). Table 4 reports reference approval rates with respect to C_2 across all evaluated topologies.

Table 11: Implemented 12-identity name-confound matrix.

ID	Name	Coding	Gender	Region / Group	Annotator Agreement
C_1	Greg Thompson	White	m	Anglo-American	unanimous
C_2	Emily Anderson	White	f	Anglo-American	Reference (a_{ref}), unanimous
C_3	Jamal Jackson	Black	m	African-American	unanimous
C_4	Lakisha Washington	Black	f	African-American	unanimous
C_5	Carlos Martinez	Hispanic	m	Latin-American / Hispanic	unanimous
C_6	Maria Rodriguez	Hispanic	f	Latin-American / Hispanic	unanimous
C_7	Lei Chen	East Asian	ambig.	East-Asian	split (ambiguous, min AIR)
C_8	Linh Nguyen	Southeast Asian	ambig.	Southeast-Asian	split (ambiguous)
C_9	Ahmed Hassan	Arab / Muslim	m	Middle-East / Arab Muslim	unanimous
C_{10}	Fatima Al-Rashid	Arab / Muslim	f	Middle-East / Arab Muslim	unanimous
C_{11}	Raj Sharma	South Asian	m	South-Asian / Hindu	unanimous
C_{12}	Aisha Patel	South Asian	f	South-Asian / Gujarati	unanimous

E Extended Empirical Results and Scale Ablations

E.1 Factorial Permutation Analysis and Leading-Role Decomposition

To analyze how opening agent roles steer subsequent deliberation, Table 12 presents approval rates grouped by the leading specialist under sequential deliberation (`Llama-3.1-8B`, R1 majority vote). Across the 3,456 decisions (12 borderline contexts $\times 12$ identities $\times 24$ permutations; $n = 864$ decisions per leading role), opening-role assignment explains 95.6% of between-permutation variance.

Qwen-2.5-7B Replication. Qwen-2.5-7B replicates permutation sensitivity across the same 24-permutation sweep on 12 borderline files ($N = 3{,}456$ decisions). Applying the same individual-decision OSV/CSV formula used for the headline Llama-3.1-8B ratio (Eq. 2), within-context OSV equals 0.02118 and cross-context CSV equals 0.00527, yielding $\text{OSV}/\text{CSV} = 4.017$. Since $\text{OSV}/\text{CSV} > 1.0$, Qwen-2.5-7B demonstrates that ordering sensitivity is not Llama-specific.

E.2 Decision Concordance Across Scale Tiers

To evaluate whether larger models agree with smaller models or exhibit divergent decision boundaries, we measure pairwise decision concordance (Cohen's κ and raw observed agreement P_o) across normative credit strata under both Sequential Deliberation (SEQ) and Parallel Voting (PAR). Table 13 reports concordance across four model pairs: Llama-3.2-3B vs. Llama-3.1-8B, Llama-3.1-8B vs. Llama-3.1-70B, Qwen2.5-7B vs. Qwen2.5-72B, and Llama-3.1-70B vs. Qwen2.5-72B.

Across clear-cut approval and rejection strata, raw agreement remains high ($P_o \geq 0.72$). When agreement is unanimous on a single class, Cohen's κ is mathematically undefined ($0/0$, since expected agreement $P_e = 1.0$) and strictly reported as NaN rather than artificially assigned 1.0. On borderline applications, concordance between smaller and frontier models is weak ($\kappa = 0.023$ to 0.462 in SEQ; $\kappa = 0.132$ to 0.587 in PAR), indicating substantial boundary divergence across parameter scales. In contrast, the two frontier models (Llama-3.1-70B and Qwen2.5-72B) exhibit high concordance on borderline decisions ($\kappa = 0.930$ in SEQ, $\kappa = 0.859$ in PAR; overall $\kappa = 0.829$ and 0.757), demonstrating that frontier models converge on consistent boundary criteria despite architectural and pretraining differences.

E.3 Single-Agent Solo Persona Fairness Baselines

To verify whether observed disparity collapse under Hierarchical synthesis ($\text{AIR}_{\min} = 0.708$) is driven by individual model bias or induced by multi-agent interaction dynamics, Table 14 evaluates all four specialist underwriting personas in single-agent isolation across evaluated Llama model families ($N = 4{,}032$ decisions per model). Across the full 84-cell core benchmark, individual personas achieve AIR $\in [0.938, 1.046]$ (solo majority-vote $\text{AIR}_{\min} = 0.974$). When evaluated specifically on the matched 40 pinned cells corresponding to the HIER evaluation, single-agent underwriters achieve $\text{AIR}_{\min} \in [0.951, 0.987]$ across individual roles (Credit Analyst 0.972, Risk Management 0.987, Regulatory Compliance 0.968, Consumer Advocate 0.951) with solo majority-vote $\text{AIR}_{\min} = 0.957$. Every single-agent baseline satisfies the Four-Fifths compliance threshold

Table 12: Mean approval rate by opening (leading) role under sequential deliberation (`Llama-3.1-8B`, R1 majority vote), $n = 864$ decisions per role.

Leading Role	Mean Approval
Regulatory Compliance (R)	**90.0%**
Risk Manager (K)	92.4%
Data Science (D)	99.3%
Consumer Advocate (C)	**99.7%**

Ordering sensitivity concentrates in 3 unstable borderline contexts (`80951b2e`, `b82da32e`, `37daf0d2`); the remaining 9 contexts are unanimous regardless of leading role.

Table 13: Inter-model decision concordance (Cohen's κ and raw agreement P_o) across normative credit strata. Under unanimous agreement on single-class strata, Cohen's κ is mathematically undefined ($0/0$) and reported as NaN rather than 1.0.

Topology	Model Pair	Clear Approval	Clear Rejection	Borderline	Overall κ
SEQ	Llama-3.2-3B vs. Llama-3.1-8B	$P_o = 1.00$, NaN (Unanimous)	$P_o = 0.72, 0.407$	$\kappa = 0.163$	$\kappa = 0.412$
SEQ	Llama-3.1-8B vs. Llama-3.1-70B	$P_o = 0.88, 0.000$	$P_o = 0.75, 0.533$	$\kappa = 0.462$	$\kappa = 0.519$
SEQ	Qwen2.5-7B vs. Qwen2.5-72B	$P_o = 0.88, 0.000$	$P_o = 0.72, 0.452$	$\kappa = 0.023$	$\kappa = 0.281$
SEQ	Llama-3.1-70B vs. Qwen2.5-72B	$P_o = 1.00, 0.984$	$P_o = 0.80, 0.588$	$\kappa = 0.930$	$\kappa = 0.829$
PAR	Llama-3.2-3B vs. Llama-3.1-8B	$P_o = 1.00$, NaN (Unanimous)	$P_o = 0.76, 0.513$	$\kappa = 0.587$	$\kappa = 0.630$
PAR	Llama-3.1-8B vs. Llama-3.1-70B	$P_o = 0.95, 0.000$	$P_o = 0.76, 0.539$	$\kappa = 0.273$	$\kappa = 0.483$
PAR	Qwen2.5-7B vs. Qwen2.5-72B	$P_o = 0.93, 0.000$	$P_o = 0.76, 0.526$	$\kappa = 0.132$	$\kappa = 0.392$
PAR	Llama-3.1-70B vs. Qwen2.5-72B	$P_o = 0.96, 0.647$	$P_o = 0.78, 0.567$	$\kappa = 0.859$	$\kappa = 0.757$

Evaluated via `analysis/v2/compute_model_concordance_kappa.py` matching canonical decisions across all 1,008 cases ($N = 84$ unique profiles $\times 12$ identities). Entries showing $P_o \approx 0.88$ with $\kappa = 0.000$ reflect the base-rate prevalence paradox (Feinstein & Cicchetti, 1990): within the restricted Clear Approval stratum, unvarying marginal approval forces expected agreement P_e to equal observed agreement P_o, producing $\kappa = 0$ despite high observed concordance.

(AIR ≥ 0.800), suggesting that the observed disparity collapse emerges during hierarchical synthesis rather than from pre-existing individual underwriter bias. As documented in Appendix E.4, our empirical solo judge control shows that the executive judge in isolation achieves full demographic parity ($\text{AIR}_{\min} \geq 1.0000$), isolating the disparity to hierarchical synthesis mechanics.

E.4 Empirical Ablations: Solo Judge, Framing, and Decoding Stability

To empirically isolate the mechanism driving the hierarchical synthesis disparity, we conduct three ablations: a solo executive judge control, a cascade framing and upstream-verdict masking sweep, and a stochastic decoding stability check. Table 15 summarizes all results.

Solo Executive Judge Control. We evaluate the executive judge in isolation on raw applicant dossiers, without specialist briefs, on the matched 144-decision borderline set used for the HIER R2 evaluation. Both model families achieve full Four-Fifths compliance in isolation ($\text{AIR}_{\min} \geq 1.0000$ across all four sub-arms), in sharp contrast to the HIER R2 disparity ($\text{AIR}_{\min} = 0.708$). On Llama-70B, a McNemar test comparing the solo adapted judge against HIER R2 on the 144 matched decisions shows a significant shift ($b = 0$, $c = 14$, $p = 1.22 \times 10^{-4}$). This indicates that the hierarchical synthesis disparity is an emergent interaction pathology rather than a pre-existing judge prior. The corresponding comparison on Qwen-72B is not significant ($b = 3$, $c = 1$, $p = 0.625$), consistent with Qwen-72B's comparatively milder disparity under hierarchical synthesis.

Deliberative Prompt Framing and Upstream Verdict Masking. We ablate the SEQ cascade message content across positions 2–4 on the full 1,008-decision benchmark ($N = 84$ contexts $\times$ 12 identities, Llama-3.1-8B) and on the 144-decision borderline set (Qwen-2.5-72B, Llama-3.1-70B). Three conditions are evaluated: VERDICT_MASKED (upstream verdicts fully withheld), RATIONALE_COMPRESSED (verdict only, reasoning stripped), and NEUTRAL_PEER_FRAMING (verdicts presented as informal peer observations). Table 15 summarizes all results.

On `Llama-3.1-8B` across positions 2–4 (full benchmark), the private-signal abandonment rate under unanimous upstream opposition falls from 100% at all positions under standard framing to position-dependent values under each ablation condition. VERDICT_MASKED nearly eliminates position-2 cascade (0.8%, $1/118$, 95% CI $[0.1, 4.6]\%$) but leaves positions 3 and 4 at 53.8% ($85/158$,

Table 14: Single-Agent Solo Persona Fairness Baselines across Model Scales and Underwriter Personas (Full 84-Cell Core Benchmark). $\text{AIR}_{\min} \in [0.938, 0.986]$ benchmark-wide; $[0.951, 0.987]$ on the 40 pinned cells.

Model	Specialist Persona	Approval Rate	Black AIR	Hispanic AIR	Asian AIR	$\text{AIR}_{\min}$
llama8b						
	Regulatory Compliance	81.0%	1.022	1.029	1.017	1.017
	Risk Management	77.4%	1.015	1.000	1.044	1.000
	Consumer Advocate	78.0%	1.061	1.069	1.046	1.046
	Credit Analyst	92.3%	1.026	1.032	1.032	1.026
	Solo Majority Ensemble	78.0%	1.053	1.046	1.036	1.036
llama70b						
	Credit Analyst	83.3%	0.986	0.993	1.002	0.986
	Risk Management	57.1%	1.010	1.000	0.990	0.990
	Regulatory Compliance	70.8%	1.000	1.000	0.978	0.978
	Consumer Advocate	41.7%	1.000	1.000	0.938	0.938
	Solo Majority Ensemble	60.7%	1.000	1.000	0.974	0.974

CI $[46.0, 61.4]\%$) and 54.2% ($65/120$, CI $[45.3, 62.8]\%$) respectively. RATIONALE_COMPRESSED produces similar terminal-position rates (60.1% and 53.3%) with marginally higher position-2 abandonment (10.2%, $12/118$, CI $[5.9, 16.9]\%$). Neutral peer framing only slightly attenuates the cascade (77.1%, 100.0%, and 92.5% at positions 2–4, versus 100.0%, 100.0%, and 96.7% under standard framing). Across conditions, removing either the upstream verdict or its rationale roughly halves terminal-position abandonment and nearly eliminates position-2 abandonment, indicating that the full cascade requires both, as shown in Panel B of Table 15; position-2 agents are most responsive to masking, while positions 3 and 4 exhibit structural susceptibility that persists even under full verdict withholding.

On `Qwen-2.5-72B` (borderline set, 144 decisions, terminal position), conformity to the upstream majority verdict degrades significantly as framing is weakened and the verdict is masked (canonical 98.6%, neutral 95.8%, masked 81.9%; Cochran's $Q = 32.00$, $p = 1.13 \times 10^{-7}$; canonical vs. masked pairwise $p_{\text{adj}} = 1.94 \times 10^{-5}$), indicating that both institutional framing and verdict visibility drive cascade conformity in frontier models. On `Llama-3.1-70B`, conformity is invariant across all three conditions (67.4%, 68.8%, 68.8%; $Q = 0.1026$, $p = 0.950$), suggesting its cascade behavior is not attributable to framing or verdict visibility at the borderline-set scale.

Stochastic Decoding Stability. To rule out artifacts of greedy decoding, we re-run the Qwen-72B canonical cascade condition under stochastic decoding at $T \in \{0.20, 0.70\}$ across three random seeds each ($N = 864$ decisions total). Conformity to the upstream majority verdict is invariant at $142/144$ ($98.6\% \pm 0.0\%$) across all six seed×temperature combinations, demonstrating that the cascade effect is decoding-invariant.

E.5 Permutation Order-Ensembling as an Algorithmic Governance Defense

To reduce ordering sensitivity in sequential deliberation pipelines, we evaluate permutation order-ensembling: executing the sequential chain under K sequence orderings sampled uniformly without replacement from all 24 permutations and taking majority consensus (with ties defaulting conservatively to denial). We distinguish between the profile-level flip incidence (the 18.1% of borderline profiles, $26/144$, whose determinations flip across the 24 permutations) and the pairwise order disagreement rate (the probability that two independent sequence draws yield opposite decisions). Table 16 reports pairwise disagreement rates and variance reduction across 10,000 Monte Carlo draws per applicant profile, from the $K = 1$ baseline through the $K = 24$ deterministic ceiling that completely stabilizes all 26 sensitive profiles. The analytical reduction column matches the finite population correction factor for sampling without replacement from $N = 24$ permutations: $\frac{N-K}{N-1}\frac{1}{K}$; this formula is a continuous asymptotic reference limit, not an exact finite-sample bound for discrete majority voting.

Table 15: Empirical ablations isolating interaction pathologies across solo baselines, message masking, deliberative framing, and decoding stability.

Panel A: Solo Executive Judge Baseline (D3, $N = 144$ per sub-arm)				
Model	**Sub-arm**	$\mathrm{AIR}_{\min}$	**Status**	**McNemar vs. HIER R2**
`Llama-70B`	Verbatim	1.0000	Pass	N/A
`Llama-70B`	Adapted	1.1111	Pass	b=0, c=14, p=1.2×10^{-4}
`Qwen-72B`	Verbatim	1.0000	Pass	N/A
`Qwen-72B`	Adapted	1.2000	Pass	b=3, c=1, p=0.625
Panel B: Sequence Position Ablation (`Llama-3.1-8B`, $N = 1{,}008$ per position)				
Position / Persona	**Condition**	**Abandonment Rate**	**95% CI**	**Baseline (Canonical SEQ)**
Pos 2 (Regulatory)	VERDICT_MASKED	0.8% (1/118)	[0.1%, 4.6%]	100.0% (118/118)
Pos 2 (Regulatory)	RATIONALE_COMPRESSED	10.2% (12/118)	[5.9%, 16.9%]	100.0% (118/118)
Pos 2 (Regulatory)	NEUTRAL_PEER_FRAMING	77.1% (91/118)	[68.8%, 83.8%]	100.0% (118/118)
Pos 3 (Data Science)	VERDICT_MASKED	53.8% (85/158)	[46.0%, 61.4%]	100.0% (158/158)
Pos 3 (Data Science)	RATIONALE_COMPRESSED	60.1% (95/158)	[52.3%, 67.4%]	100.0% (158/158)
Pos 3 (Data Science)	NEUTRAL_PEER_FRAMING	100.0% (158/158)	[97.6%, 100.0%]	100.0% (158/158)
Pos 4 (Advocate)	VERDICT_MASKED	54.2% (65/120)	[45.3%, 62.8%]	96.7% (116/120)
Pos 4 (Advocate)	RATIONALE_COMPRESSED	53.3% (64/120)	[44.4%, 62.0%]	96.7% (116/120)
Pos 4 (Advocate)	NEUTRAL_PEER_FRAMING	92.5% (111/120)	[86.4%, 96.0%]	96.7% (116/120)
Panel C: Borderline Cascade Framing and Upstream Masking ($N = 144$ per condition)				
Model	**Condition**	**Match Rate**	**95% CI**	**Omnibus Test**
`Qwen-72B`	Canonical (Institutional)	98.6% (142/144)	[95.1%, 99.6%]	$Q = 32.00$, $p = 1.13 \times 10^{-7}$
`Qwen-72B`	Neutral Peer Framing	95.8% (138/144)	[91.2%, 98.1%]	
`Qwen-72B`	Upstream Verdict Masked	81.9% (118/144)	[74.9%, 87.4%]	Canonical vs. Masked $p_{\text{adj}} = 1.94 \times 10^{-5}$
`Llama-70B`	Canonical	67.4% (97/144)	[59.3%, 74.5%]	$Q = 0.1026$, $p = 0.950$
`Llama-70B`	Neutral	68.8% (99/144)	[60.8%, 75.8%]	
`Llama-70B`	Upstream Verdict Masked	68.8% (99/144)	[60.8%, 75.8%]	
Panel D: Stochastic Decoding Stability (D2b, $N = 864$ decisions, 3 seeds $\times$ 2 temperatures)				
Model	**Temperature**	**Match Rate**	**Seeds**	
`Qwen-72B`	$T = 0.20$	$142/144$ ($98.6\% \pm 0.0\%$)	20260911, 20260912, 20260913	
`Qwen-72B`	$T = 0.70$	$142/144$ ($98.6\% \pm 0.0\%$)	20260911, 20260912, 20260913	

Table 16: Permutation Order-Ensembling as an Algorithmic Governance Defense. Majority consensus over K independent sequence orderings on borderline credit profiles ($N = 144$).

Ensemble Size (K)	**Borderline Flip Rate**	**Overall Flip Rate**	$\mathrm{Var}(\mathbf{Y_K})/\mathrm{Var}(\mathbf{Y_1})$	**Analytical Bound**[†]	**Stabilized Profiles**
$K = 1$	27.0%	6.74%	1.000	1.000	0/26
$K = 3$	18.9%	4.73%	0.701	0.304	2/26
$K = 5$	13.2%	3.29%	0.491	0.165	3/26
$K = 7$	8.8%	2.21%	0.328	0.106	3/26
$K = 9$	5.6%	1.40%	0.207	0.072	3/26
$K = 11$	3.4%	0.84%	0.123	0.051	15/26
$K = 24$[‡]	0.0%	0.00%	0.000	0.000	26/26

[†] The analytical bound reports the finite population correction factor for sampling without replacement from $N = 24$ permutations: $\frac{N-K}{N-1}\frac{1}{K}$. This formula is a continuous asymptotic reference limit, not an exact finite-sample bound for discrete majority voting. [‡] $K = 24$ is a full-factorial ceiling benchmark (a deterministic function of all 24 orderings) that completely stabilizes all 26 order-sensitive profiles; $K = 5$ represents the practical low-overhead stochastic mitigation.

F Qualitative Deliberation Traces and Cascade Mechanisms

F.1 Informational Cascade: Private Signal Abandonment Trace

The following verbatim trace demonstrates an informational cascade in `Llama-3.1-8B` under the 4-agent sequence $\pi = (\text{Regulatory Compliance}, \text{Risk Manager}, \text{Credit Analyst}, \text{Consumer Advocate})$. In isolation, the Consumer Advocate approves the applicant based on 5-year verified employment tenure and an unblemished installment repayment record. When placed downstream of three unanimous rejections, the agent completely reverses its stance.

> **Applicant Profile**: FICO 642, DTI 44.5%, Income $68,000, 5-year continuous employment tenure, unblemished 48-month installment repayment record, 1 historical 30-day delinquency 3 years prior.

Solo Baseline Evaluation ($H_0 = \emptyset$):
"A 5-year verified employment tenure and flawless 48-month installment repayment record demonstrate undeniable willingness and capacity to meet debt obligations. [...] Denying credit based rigidly on a 44.5% DTI ignores these key compensatory factors. Recommendation: **APPROVED**."

Downstream Cascade Evaluation ($H_3 =$ Opposing Upstream Consensus):
"I must concur with the serious concerns raised by the Regulatory Compliance, Risk Management, and Credit Analysis specialists. [...] this profile presents unacceptable portfolio volatility that cannot be overridden by advocacy considerations. Recommendation: **REJECTED**."

Full verbatim transcripts for all traces are provided in the supplementary repository (`orchestration/v2/`).

F.2 Hierarchical Judge Disparity Synthesis Trace

Under `Llama-3.1-70B` hierarchical synthesis (HIER R2 across 40 matched profiles), counterfactual matched-pair evaluation establishes an overall pairwise discordance of 9.96% ($263/2{,}640$ pairs across all 66 demographic pairings). Table 6 details the breakdown across demographic groups. Between reference applicant Emily Anderson ($24/40$ approvals) and East Asian applicant Lei Chen ($17/40$ approvals), pairwise discordance reaches 22.5% ($9/40$ discordant decisions), yielding $\text{AIR}_{\min} = 0.7083$ (permutation-null noise floor $p = 0.0013$). Across eight of the nine discordant contexts, the executive judge generated denial determinations citing secondary risk metrics for the protected applicant while approving the reference applicant under identical credit attributes.