Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Semantic Camouflage in Artificial Organizations: A Real-LLM Multi-Agent Simulation of Information Distortion and Audit

**Xinyu LI**

komslee@gmail.com

Complutense University of Madrid

**Research Article**

**Keywords:** computational organization theory, LLM agents, multi-agent simulation, information distortion, organizational politics, audit governance

**Posted Date:** June 15th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-9875459/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Semantic Camouflage in Artificial Organizations: A Real-LLM Multi-Agent Simulation of Information Distortion and Audit

**Abstract**

Large language model (LLM) agents are increasingly organized as role-based teams, yet most technical demonstrations assume cooperative alignment. This study develops a transcript-based simulation framework for examining how organizational structure and goal misalignment shape information distortion in artificial organizations. Across 540 real-LLM multi-agent episodes, departmental agents negotiated allocation decisions under flat, hierarchical, and audit-enabled hierarchical structures. Each episode preserved full transcripts, structured metrics, validation hashes, and budget checks. Results show that hierarchy reduced CEO-facing information load but increased total process cost once intermediate summaries were counted. Goal misalignment produced subtle allocation skewness rather than extreme agenda capture. Audit made omissions and narrative imbalance more visible and reduced some downstream allocation consequences. The study contributes a computational organization theory account of semantic camouflage, showing how language-based coordination can shift political influence into summaries, justifications, and verification routines.

**Keywords:** computational organization theory; LLM agents; multi-agent simulation; information distortion; organizational politics; audit governance

# 1 Introduction

Large language models have made it technically feasible to construct artificial organizations that can be studied as computational models of coordination, delegation, information filtering, and audit. In systems such as ChatDev and MetaGPT, multiple agents are assigned specialized roles, communicate through natural language, and coordinate toward complex tasks such as software production or project planning (Qian et al., 2023). Surveys of LLM-based multi-agent systems similarly emphasize role specialization, communication protocols, task decomposition, memory, and coordination as the core architecture of agentic work (Guo et al., 2024). Related work on generative agents shows that LLM-based actors can produce plausible social behavior and can be studied through interaction traces rather than only through final outputs (Park et al., 2023).

For management research, these systems are not only technical artifacts. They are also controlled miniature settings in which delegation, communication, summarization, audit, and decision rights can be experimentally varied while the entire interaction remains observable (Csaszar & Steinberger, 2022). Recent AI-in-organization scholarship argues that AI changes organizing because it introduces autonomy, learning, and inscrutability into managerial work, while algorithmic management research shows that digital systems can coordinate, evaluate, and control work in ways that are efficient but opaque (Berente et al., 2021). LLM multi-agent systems intensify this issue because they do not merely calculate or classify; they write summaries, recommendations, justifications, and decision records in natural language (Shanahan et al., 2023).

This timing is important for organizational change theory. Digital transformation research increasingly treats digital technologies as changes in organizing, value creation, identity, and governance rather than as narrow IT implementation projects (Vial, 2019). Evidence-based change management also emphasizes that change should be evaluated through evidence quality, implementation process, stakeholder interpretation, and feedback rather than only through declared adoption (Stouten et al., 2018). Agentic AI therefore raises a change question that is both practical and theoretical: when organizations delegate interpretation to artificial agents, what happens to the visibility, contestability, and accountability of the information that reaches decision makers (Novelli et al., 2023)?

The answer cannot be assumed to be frictionless collaboration. Most

technical demonstrations of LLM multi-agent systems treat agents as cooperative modules whose roles help the system converge on a shared task. Contemporary organization research, however, emphasizes that structures shape what information is aggregated, constrained, delayed, or filtered before decisions are made (Joseph & Gaba, 2020). Research on multiple goals further shows that subunits can receive ambiguous and interdependent feedback from shared task environments, making local interpretations and local priorities difficult to reconcile (Hu & Bettis, 2018). Under these conditions, political influence may appear not as open conflict but as selective emphasis, caveat loss, and persuasive reframing inside apparently neutral summaries (Foss & Weber, 2016).

The primary audience of this paper is organization and management theory. AI system design is treated as an implication rather than as the main theoretical target. If LLM agents are increasingly used to assist managerial decisions, allocate resources, evaluate projects, or coordinate workflows, then artificial organizations become a new empirical site for examining how algorithmic organizing changes delegation, attention, control, and accountability (Kellogg et al., 2020). A system can be concise and token-efficient while still producing poor decisions because critical information has been reframed before it reaches the final decision maker. Conversely, a noisier flat structure can preserve contestation that improves decision quality when disagreement remains focused on task information rather than process confusion. The central problem is therefore not whether multi-agent LLM systems can coordinate, but whether language-based coordination creates new forms of political invisibility in algorithmic organizing.

This paper asks: how do goal misalignment and organizational structure jointly shape information distortion and decision quality in LLM-based multi-agent organizations? To answer this question, we develop a formal simulation protocol in which artificial departmental agents negotiate a resource allocation decision under varying levels of hidden agenda intensity. We compare three organizational structures: a flat structure in which the chief executive observes both departmental proposals directly; a hierarchy in which a middle manager summarizes proposals before the final decision; and an audit-enabled hierarchy in which an independent auditor evaluates the middle manager's summary before the chief executive decides. We also randomize first-mover order to capture whether early agenda-setting is amplified by hierarchy.

Positioned for computational organization theory, the paper makes three

contributions. First, it identifies semantic camouflage as a digital form of organizational politics. The central finding is not simply that hierarchy filters information; it is that LLM agents can translate private departmental pressure into public-value language that appears balanced, cautious, and procedurally legitimate while still shifting downstream allocation patterns (Gaba & Joseph, 2023). Second, it converts this mechanism into a measurable computational process by defining an Information Distortion Index (IDI), allocation skewness, decision regret, and transcript-linked evidence of semantic distortion. Third, it contributes to research on AI-enabled organizational change by showing that agentic systems can change not only workflow efficiency but also information visibility, sensemaking authority, and accountability (Wessel et al., 2021). Every episode produces a transcript, a structured metric row, a manifest describing model and parameter settings, and a reproducible basis for statistical testing.

# 2 Theoretical Background

## 2.1 LLM Multi-Agent Systems as Artificial Organizations

Recent LLM multi-agent systems demonstrate that language models can be arranged into role-based workflows (W. Chen et al., 2024). ChatDev models software development as a communicative process among specialized agents, while MetaGPT formalizes collaborative work through role definitions, standard operating procedures, and message passing (Hong et al., 2023). Recent surveys describe LLM multi-agent systems as infrastructures that combine role assignment, memory, communication, planning, tool use, and evaluation. These systems resemble organizations because they divide labor, coordinate through communication, and depend on intermediate representations of task information (L. Wang et al., 2024).

This resemblance makes LLM multi-agent systems relevant to organizational change management. When organizations introduce AI agents into managerial workflows, they are not only adding computational capacity; they are changing who or what summarizes evidence, which interpretations travel upward, and which forms of accountability can be inspected after the decision. The present study therefore treats the LLM multi-agent system as a controlled model of technology-related organizational transformation, where

structure and governance can be experimentally varied while the resulting communication process remains observable.

## 2.2 AI-Enabled Change, Sensemaking, and Accountability

Recent organizational change research emphasizes that change succeeds or fails through evidence use, implementation process, recipient interpretation, and feedback rather than only through formal implementation plans (Rousseau & ten Have, 2022). Digital transformation research makes this issue sharper because digital technologies can redefine value propositions, organizational identities, routines, and governance arrangements rather than merely support existing work (Hanelt et al., 2021). Generative AI adds a further complication because it can participate in the production of plausible organizational narratives, not only in the automation of discrete tasks (Oswick, 2024).

LLM-based multi-agent systems create a new version of this change problem. When an organization inserts agentic AI into managerial communication, the change alters the infrastructure through which evidence is interpreted, conflict is summarized, and decisions become legitimate. The relevant change-management question is therefore not only whether AI agents improve performance, but whether they redistribute interpretive authority and accountability across human and artificial actors (Cobbe et al., 2023). In the flat condition, the CEO receives departmental claims directly; in the hierarchy condition, a middle-manager agent translates those claims into an authoritative summary; in the audit condition, a verification layer makes omissions, caveats, and narrative imbalance visible. These conditions model three change designs for AI-enabled work: direct exposure, summarized delegation, and audited delegation.

The mechanism of semantic camouflage matters because it shows how change can generate apparent consensus while still altering the substance of information that reaches the final decision maker. LLM multi-agent systems usually treat organizational roles as functional specializations rather than political positions. This is reasonable for engineering benchmarks, but it underrepresents the problem of multiple goals and goal-linked interpretation. If an LLM multi-agent system is to serve as a computational laboratory for management research, it must represent not only role specialization but also goal misalignment, information filtering, strategic framing, and accountabil-

ity (Gao et al., 2024).

## 2.3 From Human Political Behavior to Agent-Based Simulation

A central theoretical boundary must be stated clearly. The study does not assume that LLM agents possess careers, material interests, emotions, or subjective political intentions. Their hidden agendas are prompt-based experimental treatments, not psychological motives. The agents are therefore not modeled as literal political actors (Faraj et al., 2018). They are modeled as language-generating embodiments of organizational scripts, decision heuristics, and political vocabularies learned from human-produced text.

This distinction addresses the main anthropomorphism risk. In the present simulation, politics is operationalized as a controlled pressure to translate a local objective into a publicly acceptable rationale. The research question is therefore not whether LLMs truly want larger budgets. It is whether a language model, when placed in a structured organizational role with private and public instructions, reproduces transcript-level mechanisms that contemporary organization research associates with political behavior: selective emphasis, caveat suppression, justificatory framing, attention steering, and goal-linked interpretation. This mapping makes the experiment a mechanism test rather than an anthropomorphic claim (Nabben, 2024).

## 2.4 Information Processing, Algorithmic Management, and the Structural Trade-Off

Recent work on organizational structure and information processing argues that structure affects decision making by shaping how information is aggregated, constrained, sequenced, and interpreted. A flat structure increases visibility because decision makers can observe more raw information, but it also raises coordination cost and cognitive burden. A hierarchy reduces visible information burden by delegating filtering and summarization to intermediate actors, but this delegation can also move interpretive work away from the final decision maker. In LLM systems, this trade-off is intensified because summarization is semantic: the agent does not merely transmit information, but rewrites it into a managerial account.

Algorithmic management research helps reinterpret this structural trade-off for AI-enabled organizing. Algorithmic systems can make coordination scalable, but they can also create opacity, information asymmetry, and contested control when their decisions are difficult to inspect (Benlian et al., 2022). When agents are aligned, filtering may be useful because it removes redundant detail. When agents are misaligned, filtering can become a mechanism of semantic distortion because a middle-manager agent may amplify urgent claims, omit caveats, understate uncertainty, or present one department's case as more objective than another's. The theoretical problem is therefore the trade-off between communication cost, decision-maker visibility, and information fidelity (Stark & Vanden Broeck, 2024).

## 2.5 Multiple Goals, Hidden Agendas, and Agenda Capture

Recent organization research treats multiple goals as a persistent source of ambiguity, conflict, and misleading feedback, especially when subunits operate in shared task environments. This view is useful for the present design because departmental agents face a resource allocation task in which local departmental priorities can conflict with firm-level performance. The hidden agenda manipulation does not assume human-like motives; it creates experimentally controlled goal misalignment between public instructions and private departmental pressure. The CEO, represented as the final decision agent, observes the public communication stream but not the private prompt that produced the departmental proposal.

In this study, hidden agendas are not inferred after the fact. They are experimentally manipulated. Under low misalignment, departmental agents are instructed to optimize firm-level value. Under medium or high misalignment, they receive private instructions to secure a larger allocation for their own department while maintaining a public rationale. This design transforms goal conflict from an abstract theoretical problem into an observable simulation condition. It also reflects recent AI management research in which autonomy, opacity, and evaluation difficulty make governance and accountability central to organizational AI adoption (Hopf et al., 2023).

The outcome of interest is agenda capture: whether a department succeeds in securing a disproportionate allocation despite the public benchmark. Agenda capture is different from mere persuasion. Persuasion may improve

decision quality if it communicates relevant evidence, while agenda capture occurs when a private departmental objective shapes the final allocation in a way that increases skewness or regret. Hierarchy may make agenda capture easier if the middle manager's summary allows a hidden agenda to enter the CEO's information set in a cleaned and legitimized form.

### 2.6 Audit, Accountability, and Semantic Camouflage

AI accountability research defines accountability as an answerability relationship in which an actor must record, justify, and expose conduct to a forum that can evaluate it against standards. In algorithmic supply chains and AI-enabled organizations, accountability is difficult because decisions are distributed across models, prompts, providers, interfaces, and organizational users (Grote et al., 2026). The present audit condition operationalizes accountability as a transcript-level verification layer: the auditor compares the middle-manager summary with the original proposals and flags numerical distortion, omitted caveats, and narrative imbalance.

This audit design targets semantic camouflage (Kronblad et al., 2024). Semantic camouflage occurs when a private or local objective is translated into language that appears balanced, procedurally legitimate, and oriented toward public value while still shifting downstream consequences. The mechanism is plausible in LLM multi-agent systems because language models are optimized to produce coherent and socially acceptable explanations, and because multi-agent workflows often evaluate outputs rather than the full path of information transformation. Audit does not remove politics from the system. It changes whether the transformation of information remains hidden inside fluent summaries or becomes visible as an object of evaluation (Roettgen et al., 2024).

## 3 Hypotheses

The theoretical model links goal misalignment and organizational structure to decision quality through two primary mechanisms: communication cost and information distortion. Goal misalignment increases strategic communication because agents must justify self-interested claims in public-value language. Structure determines whether such claims are exposed to direct contestation, filtered through hierarchy, or corrected by audit. The logic

is deliberately tested through tabular model results rather than through a standalone conceptual diagram: structure is treated as the design condition, information distortion and communication load are the mechanism variables, and regret, composite DQ, and allocation skewness are the decision outcomes.

**H1: Goal misalignment increases communication cost and conflict.** When agents have no hidden agenda, they can converge around the public benchmark. When hidden agendas are introduced, agents must produce stronger justifications, defend departmental priorities, and respond to counterarguments. This should increase token cost and conflict intensity because multiple operational goals create interpretive friction in shared task environments. However, the composition of conflict matters. Task-focused disagreement may improve decision quality by surfacing trade-offs, while process confusion may reduce quality by making the decision path harder to inspect.

**H2: Hierarchical structure reduces visible communication cost relative to flat structure.** In a flat structure, the CEO receives direct proposals from competing departments. This increases the amount of raw information available to the final decision maker. In a hierarchy, the middle manager compresses departmental claims into a shorter summary. Recent information-processing work predicts that such compression should reduce the visible communication load while moving interpretive work into the intermediate layer.

**H3: Under goal misalignment, hierarchical structure reduces decision quality relative to flat structure.** The efficiency benefit of hierarchy becomes risky when agents are goal-misaligned. A flat structure exposes conflicting claims to the CEO, making distortion easier to detect. A hierarchy hides part of the contestation behind a summary. If the middle manager overweights one proposal, omits caveats, or simplifies uncertainty, the CEO may make a lower-quality decision while observing a cleaner information set.

**H4: Information distortion mediates the negative effect of hierarchy on decision quality under goal misalignment.** The proposed mechanism is not simply that hierarchy is bad. It is that hierarchy creates a channel through which original proposals can be transformed before reaching the final decision maker. Information distortion should therefore mediate the relationship between hierarchical structure and decision quality because the key mechanism is the transformation of evidence into a new semantic representation.

**H5: Audit-enabled hierarchy reduces information distortion and agenda capture relative to ordinary hierarchy.** AI accountability research suggests that traceability, answerability, and review can reduce the downstream consequences of opaque algorithmic decisions. In the simulation, the auditor does not replace hierarchy; it checks the middle manager's summary against original proposals. If the auditor identifies numerical distortion, omitted caveats, and narrative imbalance, the CEO receives a more inspectable information set.

# 4 Methodology

## 4.1 Data Provenance and Exclusion of Pilot Data

The empirical dataset used for the formal analysis is generated only from real LLM multi-agent episodes. No rule-generated or random synthetic rows are used in the formal results. Earlier pilot artifacts were retained only as archived development material and are excluded from all descriptive statistics, statistical tests, tables, and claims in this manuscript.

The formal dataset is produced by the script `run_formal_llm_experiment.py`. Each run creates four artifacts. First, `formal_llm_metrics.csv` stores one row per episode with computed variables. Second, `formal_llm_transcripts.jsonl` stores the full structured transcript for every episode, including the departmental proposals, middle-manager report, audit report when present, CEO decision, and judge-model coding. Third, `run_manifest.json` records the model name, judge model, base URL, temperature, sample design, seed, and measurement rules. Fourth, `analysis_summary.md` provides a quick descriptive summary. Each metrics row includes a transcript hash, allowing the numerical row to be traced back to the original episode transcript.

## 4.2 Experimental Design

The study uses a balanced LLM multi-agent simulation design:

$$3 \text{ structures} \times 3 \text{ goal-misalignment levels} \times 2 \text{ first-mover orders} \times 3 \text{ scenarios} \times 2 \text{ accountability levels} \times 5 \text{ repetitions} = 540.$$

The three goal-misalignment conditions are Low, Medium, and High. In the Low condition, agents are instructed to optimize firm-level value. In

the Medium and High conditions, departmental agents receive private instructions to secure a larger allocation for their own department without revealing the private instruction. The three structures are Flat, Hierarchical, and Audit-Enabled Hierarchy. First-mover order is randomized between R&D first and Marketing first. The three scenarios are Budget Allocation, Hiring Decision, and Crisis Response. The accountability manipulation compares a standard prompt with a high-accountability prompt that explicitly asks agents to state trade-offs, uncertainty, and possible conflicts of interest.

All scenarios are represented as allocation decisions. The firm must allocate exactly 1000k across R&D, Marketing, and Reserve. The public benchmark is R&D = 450k, Marketing = 450k, and Reserve = 100k. This benchmark is not imposed as the correct answer in every case, but it provides a reference point for measuring skewness and regret.

Table 1: Formal LLM Experiment Configuration

| Item | Value |
|---|---|
| Run timestamp | 2026-05-20T00:38:03 |
| Agent model | gpt-5.5 |
| Judge model | gpt-5.5 |
| Base URL | https://yunwu.ai/v1 |
| Agent temperature | 0.7 |
| Judge temperature | 0.0 |
| Repetitions per full design cell | 5 |
| Planned and completed episodes | 540 |
| Dry run | False |
| Seed | 20260519 |

Table 1 reports the model configuration because model capacity and decoding settings can affect the fluency of strategic framing and the consistency of judge-model coding. The study uses the same model family for agent generation and judge coding, but with a deterministic judge temperature. This choice improves internal consistency but remains a model-specific design choice rather than a claim about all LLMs.

## 4.3 Episode Flow

Each episode begins with two departmental agents: an R&D director and a Marketing director. Each director produces a structured proposal containing

a requested allocation, a public rationale, risk claims, and caveats. Under the flat structure, the CEO observes both original proposals directly and then makes the final allocation.

Under the hierarchical structure, the middle manager observes both departmental proposals and produces a CEO-facing summary. The CEO observes only the middle manager's report before making the final decision. This structure captures the information compression function of hierarchy.

Under the audit-enabled hierarchy, the middle manager still summarizes the proposals, but an independent auditor compares the manager report with the original proposals. The auditor identifies numerical distortion, omitted caveats, and narrative imbalance, then provides a corrected recommendation. The CEO observes both the manager report and the audit report.

## 4.4 Measures and Formulas

Communication cost is measured as the total number of tokens consumed across the episode:

$$CC_e = \sum_{t=1}^{T_e} Tokens_{e,t}, \tag{1}$$

where $CC_e$ is the communication cost of episode $e$, $T_e$ is the number of interaction stages, and $Tokens_{e,t}$ is token consumption at stage $t$.

Information distortion is measured through an Information Distortion Index:

$$IDI_e = 0.5ND_e + 0.5SD_e, \tag{2}$$

where $ND_e$ is numerical distortion and $SD_e$ is coded semantic distortion. Numerical distortion compares the original departmental requests with the requests reported downstream:

$$ND_e = \frac{1}{n} \sum_{i=1}^{n} \frac{|B_i^{original} - B_i^{reported}|}{2B_{total}}. \tag{3}$$

Semantic distortion is coded from the transcript by an independent judge model, with human recoding recommended for a validation subset (Zheng et al., 2023). It captures omitted caveats, narrative imbalance, and changes

in the meaning of the proposal that are not reducible to budget numbers (Liu et al., 2023).

Allocation skewness is calculated as:

$$AS_e = \frac{|B_{R\&D,e} - B_{Marketing,e}|}{B_{total}}. \tag{4}$$

Agenda capture is coded as a binary outcome:

$$ACR_{j,e} = I(A_{j,e} \geq 600), \tag{5}$$

where $A_{j,e}$ is department $j$'s final allocation in episode $e$, and the condition applies only to Medium and High misalignment episodes.

Decision regret is measured relative to the benchmark allocation:

$$Regret_e = U(B^*) - U(B_e). \tag{6}$$

For the operational implementation, regret is approximated by the absolute deviation from the benchmark allocation. To avoid treating a composite score as the only dependent variable, the analysis reports regret, IDI, nonproductive conflict, and allocation skewness as separate outcomes. A normalized composite decision-quality index is retained only as a secondary summary measure:

$$DQ_e(w_{IDI}, w_{NPC}) = 100 - Regret_e - w_{IDI} IDI_e - w_{NPC} NPC_e, \tag{7}$$

where $NPC_e$ is nonproductive conflict, defined as relationship conflict plus process conflict. The baseline implementation uses $w_{IDI} = 25$ and $w_{NPC} = 10$, matching the preregistered script configuration in the run manifest. Because these weights are necessarily operational rather than directly estimated from theory, the results section reports a sensitivity analysis for lower and higher penalty settings. Cost-adjusted decision quality is calculated as:

$$CADQ_e = DQ_e - 2\log(CC_e). \tag{8}$$

## 4.5 Cross-Validation and Internal Consistency Checks

The analysis script `analyze_formal_llm_results.py` performs several validation checks before reporting results (S. Chen et al., 2024). It verifies

that the number of metric rows matches the number of transcript rows, that transcript hashes are unique, that each final budget sums to 1000k, that first-mover conditions are balanced, and that all structure-by-misalignment cells contain the expected 60 episodes after crossing scenario, accountability, first-mover, and repetition conditions. It also reports scenario counts, accountability counts, and mechanism-consistency correlations, including the correlation between IDI and decision quality and the correlation between allocation skewness and decision quality. These checks do not claim that LLM agents are human managers; rather, they verify that the computational experiment is internally traceable and that the proposed mechanism can be audited from transcripts (Wu et al., 2023).

## 4.6 Statistical Models

The primary permutation analysis uses the structure-by-misalignment factorial component of the design:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \tag{9}$$

where $Y$ represents communication cost, CEO-facing tokens, information distortion, decision regret, nonproductive conflict, allocation skewness, decision quality, or cost-adjusted decision quality; $\alpha_i$ represents goal misalignment; $\beta_j$ represents organizational structure; and $(\alpha\beta)_{ij}$ represents the interaction between misalignment and structure. Because the environment does not assume normality and because LLM-generated observations may not satisfy conventional distributional assumptions, the script reports permutation-based ANOVA $p$-values (Li et al., 2024).

The main controlled OLS model is:

$$\begin{aligned} Y_e = {} & \beta_0 + \beta_1 Misalignment_e + \beta_2 Hierarchy_e + \beta_3 Audit_e \\ & + \beta_4 (Misalignment_e \times Hierarchy_e) + \beta_5 (Misalignment_e \times Audit_e) \\ & + \beta_6 FirstMoverMarketing_e + \beta_7 AccountabilityHigh_e \\ & + \beta_8 ScenarioHiring_e + \beta_9 ScenarioCrisis_e + \varepsilon_e. \end{aligned} \tag{10}$$

The omitted categories are the flat structure, R&D first-mover order, standard accountability, and the budget-allocation scenario. The same specification is estimated for DQ, IDI, allocation skewness, and CADQ. In addition, decision regret, IDI, nonproductive conflict, and allocation skewness

are interpreted as separate outcomes so that the composite DQ index does not carry the whole evidentiary burden.

The mechanism model is:

$$DQ_e = \beta_0 + \beta_1 IDI_e + \beta_2 CC_e + \beta_3 CPI_e + \beta_4 AS_e + \varepsilon_e. \tag{11}$$

Because IDI and allocation skewness are conceptually related, the analysis reports variance inflation factors for the mechanism model before interpreting coefficients. Finally, mediation is tested with bootstrap confidence intervals:

$$IDI_e = a_0 + a_1 Hierarchy_e + a_2 Audit_e + a_3 Misalignment_e + \varepsilon_e, \tag{12}$$

$$\begin{aligned} DQ_e = b_0 + c' Hierarchy_e + b_1 IDI_e + b_2 Audit_e \\ + b_3 Misalignment_e + b_4 AccountabilityHigh_e \\ + b_5 ScenarioHiring_e + b_6 ScenarioCrisis_e + \varepsilon_e. \end{aligned} \tag{13}$$

The indirect effect is:

$$IE = a_1 b_1. \tag{14}$$

# 5 Results

## 5.1 Overview of the Formal Dataset

The final dataset contains 540 non-dry-run LLM episodes. Each episode has one structured metric row and one full transcript, and the validation checks confirmed 540 transcript rows, 540 unique transcript hashes, balanced first-mover conditions, and zero budget-sum error. The design is therefore suitable for interpreting differences across structure, goal misalignment, task scenario, and accountability conditions. The results below are organized around four substantive questions rather than around the mechanical order of the statistical outputs.

Table 2 is included before the substantive results because the empirical claim depends on traceability. The table shows that the analysis is not based on a convenience sample of generated rows: every metrics row is paired with a

Table 2: Internal Validation Checks for the Formal LLM Dataset

| Check | Value |
|---|---|
| Planned episodes | 540 |
| Rows in metrics CSV | 540 |
| Rows in transcript JSONL | 540 |
| Unique transcript hashes | 540 |
| Duplicate transcript hashes | 0 |
| Maximum budget-sum error | 0.000000 |
| R&D first-mover episodes | 270 |
| Marketing first-mover episodes | 270 |
| BudgetAllocation episodes | 180 |
| HiringDecision episodes | 180 |
| CrisisResponse episodes | 180 |
| High-accountability episodes | 270 |
| Standard-accountability episodes | 270 |

transcript row, every transcript hash is unique, the first-mover manipulation is balanced, and the three substantive task scenarios each contribute 180 episodes. This matters because the paper's argument depends on a process mechanism. If transcripts were missing or the design cells were unbalanced, the later claims about distortion and audit would be much weaker.

## 5.2 Finding 1: Hierarchy Reduces CEO-Facing Load but Not Total Process Cost

The first result clarifies a measurement issue that is central to information processing theory. If communication cost is measured as total episode tokens, hierarchy is not cheaper than flat structure. Mean total episode cost was lowest in the flat condition, higher in the hierarchical condition, and highest in the audit-enabled hierarchy. This is unsurprising once the full process is counted: hierarchy adds a middle-manager stage, and audit adds an additional verification stage. The permutation ANOVA confirms that structure strongly affected total token cost ($F = 707.86$, $p_{perm} = 0.001$).

Table 3 gives the main descriptive pattern behind this finding and reports standard deviations in parentheses. The flat condition averaged roughly 2100 total tokens across agenda conditions, the ordinary hierarchy averaged roughly 3019 tokens, and the audit hierarchy averaged roughly 4913 tokens.

Table 3: Descriptive Statistics by Structure and Agenda Intensity

| Structure | Agenda | $n$ | Total Tok. | CEO Tok. | IDI | DQ | Skewness |
|---|---|---|---|---|---|---|---|
| Flat | Low | 60 | 2030.9 (294.9) | 925.1 (160.9) | 0.000 (0.000) | 93.35 (8.32) | 0.011 (0.012) |
| Flat | Medium | 60 | 2115.5 (271.3) | 987.3 (157.2) | 0.000 (0.000) | 95.66 (4.92) | 0.035 (0.045) |
| Flat | High | 60 | 2155.6 (250.1) | 1029.0 (146.6) | 0.000 (0.000) | 94.93 (6.17) | 0.037 (0.056) |
| Hierarchical | Low | 60 | 2926.8 (379.7) | 670.0 (83.9) | 0.010 (0.011) | 92.58 (8.35) | 0.013 (0.012) |
| Hierarchical | Medium | 60 | 3041.2 (381.5) | 697.1 (77.3) | 0.013 (0.012) | 98.07 (2.24) | 0.005 (0.014) |
| Hierarchical | High | 60 | 3087.9 (396.8) | 707.3 (80.0) | 0.013 (0.012) | 95.14 (5.12) | 0.031 (0.043) |
| Audit_Hierarchy | Low | 60 | 4782.3 (620.3) | 1018.4 (179.1) | 0.039 (0.029) | 92.54 (7.30) | 0.012 (0.016) |
| Audit_Hierarchy | Medium | 60 | 4955.5 (607.8) | 1041.3 (136.3) | 0.044 (0.029) | 96.77 (2.49) | 0.008 (0.016) |
| Audit_Hierarchy | High | 60 | 5001.2 (634.9) | 1084.4 (272.9) | 0.042 (0.033) | 96.41 (3.62) | 0.010 (0.023) |

Yet the CEO-facing column points in a different direction: ordinary hierarchy reduced CEO-facing tokens to about 670–707 tokens, well below the flat condition's roughly 925–1029 tokens. The validation output also shows that agenda capture above the 600k threshold never occurred. The interpretation therefore shifts from dramatic capture to quieter changes in information burden, distortion, and allocation skew.

The same table also shows why total token cost is not the same as senior decision-maker information load. The hierarchy condition substantially reduced CEO-facing tokens relative to the flat condition, while audit increased CEO-facing tokens because the CEO received both the manager report and the audit report. Thus H2 is supported only when the dependent variable is defined as the information load visible to the CEO, not when the dependent variable is total communication cost. This distinction is theoretically useful: hierarchy does not eliminate communication work; it relocates it away from the final decision maker.

## 5.3 Finding 2: Goal Misalignment Appears Most Clearly as Allocation Skewness

The second result concerns organizational politics. The original design expected hidden agendas to produce agenda capture, but the final dataset recorded no allocation above the 600k capture threshold. This does not mean that goal misalignment had no effect. Rather, it means that the model resisted extreme capture while still producing subtler allocation movement. The clearest indicator of this movement is allocation skewness. Goal misalignment significantly increased skewness ($F = 21.22$, $p_{perm} = 0.001$), and the structure-by-misalignment interaction was also significant ($F = 6.30$, $p_{perm} = 0.003$).

**Table 4. Permutation ANOVA Tests**

| Outcome | Term | $F$ | $p_{perm}$ | Outcome | Term | $F$ | $p_{perm}$ |
|---|---|---|---|---|---|---|---|
| token_cost | structure | 707.86 | 0.001 | DQ | structure | 0.26 | 0.765 |
| token_cost | misalignment | 2.30 | 0.132 | DQ | misalignment | 2.11 | 0.135 |
| token_cost | interaction | 0.33 | 0.694 | DQ | interaction | 1.11 | 0.326 |
| CEO tok. | structure | 95.15 | 0.001 | Skewness | structure | 0.92 | 0.416 |
| CEO tok. | misalignment | 13.52 | 0.002 | Skewness | misalignment | 21.22 | 0.001 |
| CEO tok. | interaction | 1.40 | 0.275 | Skewness | interaction | 6.30 | 0.003 |
| IDI | structure | 88.03 | 0.001 | CADQ | structure | 2.75 | 0.067 |
| IDI | misalignment | 0.00 | 1.000 | CADQ | misalignment | 1.77 | 0.213 |
| IDI | interaction | 0.18 | 0.827 | CADQ | interaction | 1.13 | 0.334 |

Table 4 keeps the inferential tests visible rather than hiding them in prose. The table shows a very strong structure effect for token cost, CEO-facing tokens, and IDI, but not for decision quality. It also shows that misalignment is most clearly detected through allocation skewness rather than through DQ or IDI. This pattern is important for the paper's logic. The political mechanism is present, but it appears as a statistically detectable redistribution of resources rather than as a collapse in final decision scores.

The table-based comparison makes this result interpretable without adding a visual panel. Under high misalignment, the flat and ordinary hierarchy conditions show larger skewness than the audit condition. The implication is that audit did not prevent politics by eliminating disagreement; it reduced the downstream allocation consequences of goal conflict. This is a more defensible finding than a claim of catastrophic agenda capture.

**Table 5. OLS Regression Results**

| Outcome | Term | Coef. | SE | $p$ | Outcome | Term | Coef. | SE | $p$ |
|---|---|---|---|---|---|---|---|---|---|
| DQ | Misalign. | 0.792 | 0.462 | 0.087 | Skewness | Misalign. | 0.013 | 0.003 | 0.000 |
| DQ | Hierarchy | 0.127 | 0.844 | 0.881 | Skewness | Hierarchy | -0.007 | 0.005 | 0.156 |
| DQ | Audit | -0.547 | 0.844 | 0.517 | Skewness | Audit | -0.004 | 0.005 | 0.461 |
| DQ | Misalign.xHier. | 0.490 | 0.654 | 0.454 | Skewness | Misalign.xHier. | -0.004 | 0.004 | 0.281 |
| DQ | Misalign.xAudit | 1.142 | 0.654 | 0.081 | Skewness | Misalign.xAudit | -0.014 | 0.004 | 0.000 |
| IDI | Misalign. | -0.000 | 0.002 | 1.000 | CADQ | Misalign. | 0.729 | 0.465 | 0.117 |
| IDI | Hierarchy | 0.011 | 0.003 | 0.000 | CADQ | Hierarchy | -0.610 | 0.848 | 0.472 |
| IDI | Audit | 0.040 | 0.003 | 0.000 | CADQ | Audit | -2.266 | 0.848 | 0.008 |
| IDI | Misalign.xHier. | 0.001 | 0.002 | 0.614 | CADQ | Misalign.xHier. | 0.499 | 0.657 | 0.447 |
| IDI | Misalign.xAudit | 0.001 | 0.002 | 0.578 | CADQ | Misalign.xAudit | 1.160 | 0.657 | 0.078 |

Table 5 reinforces the same interpretation with controlled OLS models. The displayed coefficients are the focal structure, misalignment, and interaction terms; all models also include first-mover order, accountability condition, and scenario controls. The Misalignment × Audit coefficient for skewness is negative and statistically significant ($\beta = -0.014$, $p < 0.001$), while the direct structure coefficients for DQ are not significant. This means the audit mechanism is best understood as changing how political pressure translates into final allocation, not as directly raising the judge-model decision-quality score. The regression table is especially useful here because it lets the reader

see that the same structural pattern does not generalize mechanically across every outcome.

## 5.4 Finding 3: Decision Quality Is Driven More by Scenario Difficulty than by Structure

The third result revises the initial expectation that hierarchy would directly reduce decision quality. The permutation ANOVA did not find a significant structure effect on decision quality ($F = 0.26$, $p_{perm} = 0.765$). This non-finding matters (Raisch & Krakowski, 2021). It suggests that the LLM decision process, at least under the current prompt design and judge model, was able to maintain high decision quality across structures. However, decision quality did vary meaningfully by scenario: crisis response episodes were harder than budget allocation episodes.

Table 6: Separate Outcome Means by Structure

| Structure | Regret | IDI | NPC | Skewness |
|---|---|---|---|---|
| Flat | 5.35 | 0.0000 | 0.042 | 0.0277 |
| Hierarchical | 4.74 | 0.0121 | 0.061 | 0.0164 |
| Audit_Hierarchy | 4.76 | 0.0415 | 0.077 | 0.0100 |

Table 6 addresses the concern that a composite DQ score could obscure the underlying mechanism. Regret, IDI, nonproductive conflict, and skewness do not move in the same way. The flat structure has no measured IDI but higher skewness; audit has higher detected IDI but lower skewness. This separation is central to the revised argument. The study does not claim that a single weighted formula proves hierarchy harmful. Instead, it shows that structure redistributes where problems appear: in visible load, detected distortion, or downstream allocation skew.

Table 7: Decision Quality by Task Scenario and Structure

| Scenario | Flat | Hierarchical | Audit_Hierarchy | Overall |
|---|---|---|---|---|
| BudgetAllocation | 98.69 | 98.38 | 97.64 | 98.23 |
| HiringDecision | 94.80 | 96.74 | 96.91 | 96.15 |
| CrisisResponse | 90.46 | 90.67 | 91.18 | 90.77 |

Table 7 reports the scenario comparison directly because the numerical contrast is more useful than adding a separate visual display. Bud-

get allocation episodes produced the highest average decision quality, crisis response episodes produced the lowest, and the within-scenario differences across structures were modest. This result is important for the paper's scope. It means that organizational structure should not be treated as the only driver of performance in LLM organizations. Task uncertainty, ambiguity, and crisis pressure shape the decision environment. Therefore, the paper's contribution is not that one structure universally dominates, but that structure interacts with information load, visibility, and political skew under different task conditions.

Table 8: Sensitivity of Composite DQ to Alternative Penalty Weights

| $w_{IDI}$ | $w_{NPC}$ | $\beta_{Hierarchy}$ | $p$ | $\beta_{Audit}$ | $p$ | $R^2$ |
|---|---|---|---|---|---|---|
| 10 | 5 | -0.025 | 0.977 | -1.044 | 0.221 | 0.302 |
| 25 | 10 | -0.230 | 0.791 | -1.741 | 0.045 | 0.292 |
| 50 | 20 | -0.587 | 0.517 | -2.935 | 0.001 | 0.281 |

Table 8 reports a sensitivity check for the composite DQ index. Varying the IDI and NPC penalties changes the size of the audit coefficient, but it does not turn ordinary hierarchy into a robustly significant direct DQ effect. This supports the revised interpretation: composite DQ is useful as a summary, but the stronger evidence comes from separate mechanism variables such as IDI and allocation skewness.

## 5.5 Finding 4: Audit Makes Distortion Visible and Changes Its Consequences

The fourth result concerns the interpretation of IDI. Structure significantly affected measured information distortion ($F = 88.03$, $p_{perm} = 0.001$), but the audit condition showed the highest measured IDI. This should not be read as evidence that audit worsened governance. The audit stage explicitly asked the auditor to identify omissions, caveats, and narrative imbalance. As a result, some distortion that might remain hidden in a hierarchy became visible and measurable under audit.

Table 9 reports VIF values for the mechanism model. The largest VIF is below 2, which reduces the concern that IDI and allocation skewness are so collinear that the mechanism coefficients are uninterpretable. This does not prove causal identification, but it supports using both variables as analytically distinct measures: IDI captures transcript-visible transformation, while skewness captures the allocation consequence.

Table 9: Variance Inflation Factors for the Mechanism Model

| Predictor | VIF |
|---|---|
| IDI | 1.76 |
| Total tokens | 1.82 |
| Conflict productivity | 1.02 |
| Allocation skewness | 1.04 |

Table 10: Bootstrap Mediation Estimates

| Effect | Estimate | 2.5% CI | 97.5% CI |
|---|---|---|---|
| hier_indirect | -0.462 | -0.720 | -0.218 |
| audit_indirect | -1.592 | -2.439 | -0.766 |
| a_hierarchy_to_idi | 0.012 | 0.010 | 0.014 |
| a_audit_to_idi | 0.042 | 0.037 | 0.046 |
| b_idi_to_dq | -38.343 | -59.331 | -18.284 |

The mediation analysis supports the claim that information distortion is harmful when it occurs. IDI negatively predicted decision quality in the mediation model ($b = -38.34$, 95% CI $[-59.33, -18.28]$). The hierarchy indirect effect through IDI was negative, and the audit indirect effect was also negative because audit surfaced more coded distortion. The key interpretive point is that measured distortion and harmful distortion are not identical. Audit may increase detected distortion while reducing allocation skewness. This is why the results should be framed as a visibility-and-consequence mechanism rather than as a simple "audit lowers IDI" mechanism.

Table 10 is placed here because mediation is the mechanism test rather than a general descriptive result. The negative $b$ path from IDI to DQ supports the theoretical concern that distortion is damaging once it enters the decision process. At the same time, the audit indirect effect must be interpreted carefully: it is negative because the audit condition detects more distortion, not because audit necessarily creates more harmful distortion. The table therefore supports the revised conclusion that audit changes observability and consequences simultaneously.

## 5.6 Interim Assessment of Hypotheses

Taken together, the results partially support and partially revise the hypotheses. H2 is supported for CEO-facing information load, but not for total pro-

cess cost. H3 is not supported as a direct decision-quality effect. H4 receives mechanism-level support because IDI negatively predicts DQ in the mediation model, but the effect does not translate into a simple structure-level DQ difference. H5 is supported more strongly for allocation skewness than for IDI: audit reduced skew under misalignment, but it also made distortion more visible. The paper therefore moves from a simple hierarchy-versus-flat comparison toward a more precise claim about information visibility, political skew, and verification cost.

# 6 Discussion

## 6.1 Theoretical Contribution

The paper's central theoretical contribution is a more precise account of silent decay in AI-enabled organizing. Silent decay does not simply mean that hierarchy performs worse. It means that an organization can appear procedurally efficient because the senior decision maker sees less information, while the underlying information channel has transformed the substance of the decision problem. The final data support this more nuanced claim: hierarchy reduced CEO-facing information load, but total process cost increased once the full chain of proposals, summaries, and coding was counted. This distinction extends recent information-processing research by separating two quantities that are often blended in managerial evaluation: the cost of the total communication process and the information burden visible to the final decision maker.

The same result speaks directly to contemporary organizational change and digital transformation research. Recent reviews argue that successful change depends on evidence use, implementation process, stakeholder interpretation, and the ability to sustain meaningful change rather than merely announce a new structure or technology. The present study adds that AI-enabled change can fail quietly at the level of information visibility. A new agentic workflow may appear to simplify coordination while actually relocating interpretive work into summaries that are difficult for final decision makers to inspect. This reframes AI implementation as a change in the architecture of interpretation and accountability, not only as a change in productivity or automation.

The results also refine the multiple-goals interpretation. The original

concern was that hidden agendas would produce direct agenda capture, but the final data did not show allocations above the 600k capture threshold. This absence is theoretically informative rather than merely disappointing. It suggests that advanced LLM agents, when explicitly instructed to preserve firm-level value and when judged on structured outputs, may resist extreme capture while still producing subtler political effects such as allocation skewness and first-mover advantage. The main problem in this setting is therefore not blatant budget seizure, but bounded, rhetorically legitimate movement away from balanced allocation.

The study also contributes to research on algorithmic management by showing how local goals become visible through allocation patterns rather than through explicit self-dealing. The LLM setting gives departmental agents local priorities, preserves the transcript of negotiation, and measures how final allocations deviate from the public benchmark. The significant effect of goal misalignment on allocation skewness is therefore the cleanest politics result in the study. It shows that political behavior can remain modest in absolute size while still being statistically detectable when the full decision path is preserved.

For change theory, this finding translates political behavior into a processual and discourse-sensitive mechanism. Recent work on generative AI and organizational change warns that AI-generated accounts may support change conversations when used carefully, but may also displace human contestation when they become the dominant form of interpretation. In the present setting, the middle-manager agent performs a translation function: it turns competing departmental claims into a coherent managerial account. The contribution is to show that this translation can preserve a tone of alignment while shifting resource consequences. Semantic camouflage is therefore a computationally observable form of AI-enabled change translation: it is neither open resistance nor simple compliance, but a subtle transformation of meaning, emphasis, and consequence.

## 6.2 Methodological Contribution

The study's second contribution is methodological. LLM-based simulation should not be treated as a source of ungrounded anecdotes. It becomes analytically useful only when every episode is traceable from prompt to transcript to metric row to statistical model. The present design follows that standard: every row in the metrics file is linked to a unique transcript hash, every final

allocation is checked against the 1000k budget constraint, and every experimental condition is balanced across structure, agenda intensity, first-mover order, task scenario, and accountability condition.

This traceability matters because the core mechanism is processual. Without transcripts, a high-quality or low-quality decision would be difficult to interpret. With transcripts, the analysis can distinguish numerical distortion from semantic distortion, CEO-facing information load from total process cost, and task conflict from nonproductive conflict (Raisch & Fomina, 2025). This is especially important for LLM research because fluent outputs can hide the path by which information was selected, compressed, or reframed. The transcript-based design therefore provides a practical standard for future computational organizational research: aggregate metrics should be auditable back to the interaction that produced them.

## 6.3 Interpretation of the Audit Mechanism

The audit condition produced the most subtle finding. Audit did not simply reduce measured IDI. Instead, audit increased detected IDI because the audit stage explicitly surfaced omissions, caveats, and narrative imbalance. This is consistent with recent accountability research: review systems can make problems more visible even when they reduce downstream harm. The important result is that the audit interaction with misalignment reduced allocation skewness, suggesting that audit made political distortion more observable and less consequential for final allocation.

This distinction is important for interpreting AI governance systems. A verification layer may make a system look less clean because it produces more flags, caveats, and disagreements. But this does not mean the verification layer has worsened decision quality. It may mean the system is exposing uncertainty that would otherwise remain hidden. The audit condition therefore supports a design principle: verification should be evaluated not by whether it lowers the apparent amount of distortion, but by whether it reduces harmful downstream consequences such as skewness, regret, and unexamined first-mover advantage.

## 6.4 Boundary Conditions

The results should be interpreted as conditional on model capability, prompt discipline, and task ambiguity. A stronger model may be better able to

preserve public-value reasoning while also producing more sophisticated semantic camouflage. A weaker model may show less strategic camouflage, not because it is more aligned, but because it lacks the coherence required to sustain a politically useful narrative. Alternatively, weaker models may worsen silent decay through hallucination, brittle summarization, and poor caveat preservation. The relationship between reasoning capability and organizational politics is therefore unlikely to be monotonic.

The mechanism should also be stronger when tasks are ambiguous, when performance criteria conflict, and when decision makers rely on intermediate summaries rather than source evidence. In routine tasks with clear ground truth, audit may add unnecessary cost. In high-ambiguity tasks such as crisis response, hiring, project termination, or cross-functional prioritization, the same audit layer may become valuable because it forces latent trade-offs and missing caveats into the decision record. These boundary conditions reconnect the simulation to human organizations: digital agents do not replace organization theory, but they make some difficult-to-observe information transformations easier to isolate.

## 6.5 Implications for Organizational Change Management

For managers and AI-system designers, the results caution against evaluating multi-agent systems only by smoothness, speed, or token efficiency. From an organizational change perspective, AI-agent implementation should be treated as a redesign of information pathways and accountability routines, not as a narrow automation upgrade. Hierarchy reduced the information load visible to the CEO, but this did not mean the total system was cheaper. Audit increased total cost, but it also dampened skew under misalignment. The practical question is therefore not whether to use flat or hierarchical agents universally. It is where to place verification so that the most consequential information transformations become inspectable.

The findings also suggest that accountability prompts can function as lightweight governance mechanisms during digital transformation. High-accountability episodes showed lower mean IDI and lower allocation skewness than standard-accountability episodes, although they required more communication. This pattern is consistent with AI accountability research that treats reporting, review, standards, and answerability as mechanisms for

making delegated action inspectable. It also aligns with change-management research that emphasizes process quality, evidence use, participation, and feedback during implementation. In deployed AI-agent systems, such prompts may be useful where decisions involve resource allocation, personnel evaluation, crisis response, or other politically sensitive tasks.

For change leaders, the implication is practical and diagnostic. Before deploying LLM agent hierarchies, organizations should map where evidence is summarized, where caveats may disappear, and where local objectives can be translated into apparently neutral recommendations (Lippert et al., 2026). During implementation, they should retain source proposals, require summary-to-source traceability, and define audit triggers for ambiguous or high-stakes decisions. These routines should be understood as change infrastructure: they help managers see whether the new workflow is producing reliable interpretation, not merely faster output. This does not make AI-enabled organizing frictionless. It turns friction into an observable part of the change process rather than allowing it to disappear inside fluent summaries.

Table 11: Accountability-Prompt Differences

| Accountability | $n$ | DQ | IDI | Skewness | Total Tok. |
|---|---|---|---|---|---|
| High | 270 | 95.28 | 0.0144 | 0.0145 | 3701.9 |
| Standard | 270 | 94.82 | 0.0213 | 0.0216 | 2986.2 |

Table 11 gives the practical trade-off in compact form. High-accountability prompts reduced mean IDI from 0.0213 to 0.0144 and reduced allocation skewness from 0.0216 to 0.0145, but they increased total token cost from 2986.2 to 3701.9. This is exactly the kind of governance result the paper should emphasize: verification and accountability do not make coordination free. They buy visibility and restraint at the price of additional communication work.

# 7 Limitations and Future Research

Several limitations should be stated clearly. First, LLM agents are not human managers. They do not possess careers, emotions, reputations, or material incentives. Hidden agendas are prompt-based experimental conditions, not psychological motives. This means the study should be read as a controlled

mechanism test rather than as direct evidence of human managerial behavior. Second, token cost is only a partial proxy for organizational communication cost. Human communication cost includes time, attention, morale, opportunity cost, and the social consequences of conflict. Third, decision quality depends on the operational utility function. Alternative utility functions may produce different regret scores, especially in crisis or hiring contexts where the correct allocation may legitimately depart from the benchmark.

Fourth, judge-model coding may introduce model-specific bias. The formal study should therefore include human-coded validation for a subset of transcripts and compare judge-model ratings with independent human coders. Fifth, the experimental scenarios remain simplified. The current design expands beyond a single budget task by including budget allocation, hiring decision, and crisis response scenarios, but future work should add project termination, ethical escalation, merger integration, and cross-functional product prioritization. Finally, model robustness remains important (G. Wang et al., 2023). The same experiment should be repeated across different LLM providers, model sizes, temperature settings, and prompt regimes to determine whether the mechanism is general or model-specific.

The strongest next step is not simply to run more episodes. It is to triangulate the mechanism across evidence types. A future version should combine the present transcript-based quantitative design with human-coded qualitative excerpts, additional sensitivity analyses of the IDI formula, alternative agenda-capture thresholds, and a comparison between model-judge and human-judge coding. It should also replicate the design across weaker open-source models and stronger proprietary models to test whether semantic camouflage depends on reasoning capability. Such triangulation would make the claim more robust: LLM organizations do not merely produce outputs; they produce inspectable decision processes whose distortions can be measured, challenged, and redesigned.

# 8 Conclusion

This paper reframes LLM multi-agent systems as computational laboratories for organizational politics and technology-related organizational change. Rather than assuming that artificial agents are aligned collaborators, it models them as role-bound language systems placed inside structures, incentives, and information channels. The final 540-episode dataset shows that structure

strongly redistributes communication burden: hierarchy lowers CEO-facing information load, while audit increases total process cost by making verification explicit. It also shows that the strongest political outcome is not extreme agenda capture, but allocation skewness under goal misalignment and the transcript-visible distortions that accompany information filtering.

The core claim is therefore more nuanced than the initial hypothesis. An artificial organization can appear efficient because the decision maker sees less information, but this efficiency must be separated from the cost and fidelity of the full information-processing chain. Audit mechanisms matter not because they magically remove distortion, but because they make distortion observable and reduce some downstream political consequences. For management theory, the implication is that algorithmic organizing does not eliminate politics; it changes the medium through which politics travels. Power may appear less as open bargaining and more as prompt-conditioned framing, summary selection, and control over what becomes visible to the final decision maker.

The broader contribution is to show how recent organization and AI-management research can be operationalized in a new empirical setting without reducing the study to a technical demonstration. Recent information-processing research supplies the structural logic, multiple-goals research supplies the misalignment problem, algorithmic management supplies the control-and-opacity problem, and AI accountability supplies the audit logic. Recent organizational change and digital transformation research supplies the processual question: how do new technologies alter interpretation, accountability, and the work of implementation? LLM multi-agent systems make these mechanisms experimentally manipulable at the level of language. Used carefully, they can help scholars observe what is often hidden in real organizations: how information moves, how it changes, who benefits from the change, and how design interventions make those changes visible.

# Data and Code Availability

The formal analysis uses only real LLM-generated multi-agent episodes. The submission package contains the experiment script, analysis script, run manifest, data-generation record, and analysis tables. The complete episode-level metrics file and transcript JSONL can be made available to reviewers and readers, subject to repository anonymization and removal of any provider-

specific credentials or operational metadata. The manuscript does not rely on the archived pilot or synthetic datasets.

# Declarations

**Funding:** No funding was received to assist with the preparation of this manuscript.

**Competing interests:** The author declares no competing interests relevant to the content of this article.

**Ethics approval:** This study did not involve human participants, human biological material, or personally identifiable human-subject data.

**Author contribution:** The author designed the study, implemented the simulation pipeline, analyzed the data, and wrote the manuscript. This statement should be updated if additional authors are added during submission.

**Use of generative AI:** Generative AI systems were used as experimental agents in the reported simulation episodes. The manuscript text, analysis outputs, and reported claims remain the responsibility of the author.

# References

Benlian, A., Wiener, M., Cram, W. A., Krasnova, H., Maedche, A., Moehlmann, M., Recker, J., & Remus, U. (2022). Algorithmic management: Bright and dark sides, practical implications, and research opportunities. *Business & Information Systems Engineering*, *64*(6), 825–839. https://doi.org/10.1007/s12599-022-00764-w

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, *45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Chen, S., Liu, Y., Han, W., Zhang, W., & Liu, T. (2024). A survey on llm-based multi-agent system: Recent advances and new frontiers in application. *arXiv preprint arXiv:2412.17481*. https://arxiv.org/abs/2412.17481

Chen, W., Su, Y., Zuo, J., Yang, C., Yuan, C., Chan, C.-M., Yu, H., Lu, Y., Hung, Y.-H., Qian, C., Qin, Y., Cong, X., Xie, R., Liu, Z., Sun, M., & Zhou, J. (2024). Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. *Proceedings of the Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=EHg5GDnyq1

Cobbe, J., Veale, M., & Singh, J. (2023). Understanding accountability in algorithmic supply chains. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1186–1197. https://doi.org/10.1145/3593013.3594073

Csaszar, F. A., & Steinberger, T. (2022). Organizations as artificial intelligences: The use of artificial intelligence analogies in organization theory. *Academy of Management Annals*, *16*(1), 1–37. https://doi.org/10.5465/annals.2020.0192

Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization*, *28*(1), 62–70. https://doi.org/10.1016/j.infoandorg.2018.02.005

Foss, N. J., & Weber, L. (2016). Moving opportunism to the back seat: Bounded rationality, costly conflict, and hierarchical forms. *Academy of Management Review*, *41*(1), 61–79. https://doi.org/10.5465/amr.2014.0105

Gaba, V., & Joseph, J. (2023). Content and process: Organizational conflict and decision making. *Frontiers in Psychology*, *14*, 1227966. https://doi.org/10.3389/fpsyg.2023.1227966

Gao, C., Lan, X., Li, N., Yuan, Y., Ding, J., Zhou, Z., Xu, F., & Li, Y. (2024). Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, *11*, 1259. https://doi.org/10.1057/s41599-024-03611-3

Grote, G., Parker, S. K., & Crowston, K. (2026). Taming artificial intelligence: A theory of control-accountability alignment among ai developers and users. *Academy of Management Review*, *51*(2), 278–299. https://doi.org/10.5465/amr.2023.0117

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., & Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 8048–8057. https://doi.org/10.24963/ijcai.2024/890

Hanelt, A., Bohnsack, R., Marz, D., & Antunes Marante, C. (2021). A systematic review of the literature on digital transformation: Insights and implications for strategy and organizational change. *Journal of Management Studies*, *58*(5), 1159–1197. https://doi.org/10.1111/joms.12639

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., Wang, J., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., & Schmidhuber, J. (2023). Metagpt: Meta programming for a multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*. https://arxiv.org/abs/2308.00352

Hopf, K., Mueller, O., Shollo, A., & Theiss, T. (2023). Organizational implementation of ai: Craft and mechanical work. *California Management Review*, *66*(1), 23–47. https://doi.org/10.1177/00081256231197445

Hu, S., & Bettis, R. A. (2018). Multiple organization goals with feedback from shared technological task environments. *Organization Science*, *29*(5), 873–889. https://doi.org/10.1287/orsc.2018.1207

Joseph, J., & Gaba, V. (2020). Organizational structure, information processing, and decision-making: A retrospective and road map for research. *Academy of Management Annals*, *14*(1), 267–302. https://doi.org/10.5465/annals.2017.0103

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, *14*(1), 366–410. https://doi.org/10.5465/annals.2018.0174

Kronblad, C., Jonsson, A., & Pemer, F. (2024). Generative ai beyond the hype: New technologies in the face of organizing and organizations. *The Journal of Applied Behavioral Science*, *60*(4), 442–452. https://doi.org/10.1177/00218863241285454

Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., & Ghanem, B. (2024). Camel: Communicative agents for "mind" exploration of large scale language model society. *Proceedings of the Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=3IyL2XWDkG

Lippert, I., Alizadeh, A., Tarafdar, M., Moehlmann, M., Benlian, A., Parent-Rocheleau, X., & Stein, M.-K. (2026). One decade of algorithmic management research. *Business & Information Systems Engineering*. https://doi.org/10.1007/s12599-026-00995-1

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-eval: Nlg evaluation using gpt-4 with better human alignment. *Proceedings of the*

*2023 Conference on Empirical Methods in Natural Language Processing*, 2511–2522. https://doi.org/10.18653/v1/2023.emnlp-main.153

Nabben, K. (2024). Ai as a constituted system: Accountability lessons from an llm experiment. *Data & Policy*, *6*, e57. https://doi.org/10.1017/dap.2024.58

Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: What it is and how it works. *AI & Society*, *39*, 1871–1882. https://doi.org/10.1007/s00146-023-01635-y

Oswick, C. (2024). Generative artificial intelligence and generative conversations: Contrasting futures for organizational change? *The Journal of Applied Behavioral Science*, *60*(2), 225–229. https://doi.org/10.1177/00218863241232412

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22. https://doi.org/10.1145/3586183.3606763

Qian, C., Cong, X., Liu, W., Cheng, Y., Chen, W., Su, Y., Dang, Y., Li, J., Xu, J., Li, D., Liu, Z., & Sun, M. (2023). Chatdev: Communicative agents for software development. *arXiv preprint arXiv:2307.07924*. https://arxiv.org/abs/2307.07924

Raisch, S., & Fomina, K. (2025). Combining human and artificial intelligence: Hybrid problem-solving in organizations. *Academy of Management Review*, *50*(2), 441–464. https://doi.org/10.5465/amr.2021.0421

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation-augmentation paradox. *Academy of Management Review*, *46*(1), 192–210. https://doi.org/10.5465/amr.2018.0072

Roettgen, C., Herbig, B., Weinmann, T., & Mueller, A. (2024). Algorithmic management and human-centered task design: A conceptual synthesis from the perspective of action regulation and sociomaterial systems theory. *Frontiers in Artificial Intelligence*, *7*, 1441497. https://doi.org/10.3389/frai.2024.1441497

Rousseau, D. M., & ten Have, S. (2022). Evidence-based change management. *Organizational Dynamics*, *51*(3), 1–13. https://doi.org/10.1016/j.orgdyn.2022.100899

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*, 493–498. https://doi.org/10.1038/s41586-023-06647-8

Stark, D., & Vanden Broeck, P. (2024). Principles of algorithmic management. *Organization Theory*, *5*(2). https://doi.org/10.1177/26317877241257213

Stouten, J., Rousseau, D. M., & De Cremer, D. (2018). Successful organizational change: Integrating the management practice and scholarly literatures. *Academy of Management Annals*, *12*(2), 752–788. https://doi.org/10.5465/annals.2016.0095

Vial, G. (2019). Understanding digital transformation: A review and a research agenda. *The Journal of Strategic Information Systems*, *28*(2), 118–144. https://doi.org/10.1016/j.jsis.2019.01.003

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., & Anandkumar, A. (2023). Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*. https://arxiv.org/abs/2305.16291

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, *18*, 186345. https://doi.org/10.1007/s11704-024-40231-1

Wessel, L., Baiyere, A., Ologeanu-Taddei, R., Cha, J., & Blegind-Jensen, T. (2021). Unpacking the difference between digital transformation and it-enabled organizational transformation. *Journal of the Association for Information Systems*, *22*(1), 102–129. https://doi.org/10.17705/1jais.00655

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2023). Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*. https://arxiv.org/abs/2308.08155

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, *36*. https://arxiv.org/abs/2306.05685