Leadership as Coordination Control: Behavioral Signatures and the Recovery-Advantage Boundary in Multi-Agent LLM Teams

Anonymous authors
Paper under double-blind review

Abstract

Team science holds that leadership is *contingent*: it helps only under specific conditions, and capable, autonomous teams may need none at all. We ask the analogous question for multi-agent LLM teams: under what *measurable* conditions does process-level coordination control add value, and do those conditions match what team science predicts? Answering it takes measurement, not just accuracy: we use *behavioral signatures* (majority lock-in, exploration, recovery from an incorrect round-0 consensus) and *per-action ablations*, clean because each controller is an explicit action set rather than a monolithic prompt.

We operationalize three classical leadership styles (transactional, transformational, situational) as controllers over a shared action vocabulary (`explore`, `revise`, `accept`, `synthesize`), holding the agent set and final aggregation fixed. Team science (Hackman, 2002; Bass, 1985; Hersey & Blanchard, 1969) supplies the substrate: Bass's two-component structure maps to the `accept` and `revise` actions, giving the transactional controller its decomposition. A matched controller with the same actions but an *arbitrary* rule recovers no better than majority voting, so it is the theory-derived rule, not the vocabulary, that does the work.

Across four task regimes and three open-weight model families on a single backend, no controller dominates by accuracy, as the contingency view predicts. Against a *shared* round-0 vote, generated once and reused across conditions, transactional control matches the vote on all 12 (model, regime) combinations to within 1.3pp, and accuracy gains appear in only two of the 36 leadership entries, situational and transformational, both on the single `llama-4-scout` social combination, where the round-0 majority is unreliable. Against the stronger flat baseline, only situational still gains (+8pp). A *recovery-advantage* account, tested with four boundary probes, says when a controller beats plain interaction: only where the round-0 majority is unreliable, the task is recoverable, and undirected interaction does not already repair it. These conditions map onto contingency theory (leadership substitutes, path-goal redundancy, and the situational readiness gap), so a largely null accuracy result is what the theory predicts, not a failure of the controllers. We read process-level coordination control as a contingency to be measured and theory-mapped, not a leaderboard to be topped.

1 Introduction

Large language model (LLM) agents are increasingly deployed in collaborative settings where multiple agents deliberate, critique one another, and produce a joint answer. Recent multi-agent systems (MAS) work has explored several engineering patterns that improve over single-agent inference: multi-agent debate (Du et al., 2023; Liang et al., 2024; Chan et al., 2024), role specialization and pipelined decomposition (Hong et al., 2024; Wu et al., 2024; Li et al., 2023; Chen et al., 2024), and self-refinement or critic-then-resolve loops (Madaan et al., 2023; Shinn et al., 2023). These patterns operate at the *knowledge level*: they shape what agents reason about, how reasoning is decomposed across agents, or whether a candidate answer is refined before final aggregation. Coordination is then implicit in the choice of debate format, decomposition rule, or refinement loop. We ask a complementary, *process-level* question: given a fixed agent set and a fixed

knowledge-combination scheme, what control vocabulary determines how disagreement is raised, revised, accepted, or reopened across rounds, and how do different choices in that vocabulary shape collective behavior? Team science suggests the answer is conditional. Leadership there is not universally beneficial but *contingent* on team state: subordinate ability and clear tasks can *substitute* for it entirely (Kerr & Jermier, 1978), and a leader adds value only by supplying a function the team is not otherwise getting (Hackman & Walton, 1986). The sharper question is therefore *when*, under what measurable conditions, process-level control adds value at all, and whether those conditions match what team science predicts.

We propose two methodological primitives for studying this process-level question. First, *behavioral signatures* (majority lock-in, exploration rate, recovery from incorrect round-0 consensus, and dissent preservation, the last reflected in the recovery rate; Section 4.4) replace single-number accuracy as the primary measurement target. Final accuracy is informative but varies sharply across model and regime combinations and is sensitive to per-model stochasticity (Appendix B.6); behavioral signatures, especially majority lock-in, produce within-controller contrasts that reproduce even when absolute accuracy rankings shift. Second, *per-action ablations* expose which components of a controller carry the effect in which regime. Each controller is specified as a small, explicit set of control actions rather than as a single black-box prompt, so removing or substituting one action while keeping the rest active is a clean operation. Prompt-based controllers do not afford equally clean component-level ablations, since the mapping between prompt text and behavioral component runs through the LLM's interpretation.

As a substantive case study, we operationalize three classical leadership styles (transactional, transformational, situational) as process-level controllers over a small interaction-level action space (such as `explore`, `revise`, `accept`, `synthesize`; see Section 4.3 for the full set). Team science (Hackman, 2002; Salas et al., 2005; Bass, 1985; Hersey & Blanchard, 1969) supplies the principled substrate: a theoretical organization of process, cognitive, and affective coordination levels (we concentrate on the first), and component-decomposable action sets. Bass's two-component structure (contingent reward, management by exception) maps directly to the `accept` and `revise` actions within transactional control, and a parallel decomposition (broadcast goal, differentiated directives) applies to transformational control. The substrate supplies theory-grounded action decomposition rather than new coordination mechanisms beyond multi-agent debate, role specialization, and self-refinement already explored at the knowledge level. A controller decides not what an agent should think, but whether the team should reopen deliberation, reduce dissent, or close the round; the two layers are complementary and our process-level vocabulary composes with, rather than replaces, existing knowledge-level mechanisms.

We test these methods on four task regimes (closed-ended QA (Talmor et al., 2019; Geva et al., 2021), abductive ambiguity (Bhagavatula et al., 2019), social-norm ambiguity (Forbes et al., 2020), and a mixed-workload benchmark interleaving the three) across three open-weight model families (`gpt-oss-120b`, `llama-4-scout`, `gemma-4-31B-it`) served through a single inference backend, plus a second tier of four boundary probes (a cross-domain extension to MATH-500 Level 5 (Hendrycks et al., 2021; Lightman et al., 2024), adversarial NLI (Nie et al., 2020), Winogrande (Sakaguchi et al., 2021), and a contested moral-judgment set (Lourie et al., 2021)) that deliberately vary round-0 reliability and recoverability (Section 6). The findings answer the contingency question. No controller dominates by accuracy, as the contingency view predicts, and the conditions under which one helps are organized by a single measurable axis, the reliability of the independent round-0 majority (how often it is correct): a controller adds value only where that majority is unreliable, the task is recoverable (its incorrect majority can be repaired), and undirected interaction does not already repair it. We then map this measured boundary back to team-science contingency theory (Section 6.3), where each region corresponds to a named construct: leadership substitutes, path-goal redundancy (undirected interaction already reaches the answer, so direction adds nothing), and the situational readiness gap (the team can improve, but only with direction).

Our contributions are:

- Two measurement primitives for process-level coordination control: *behavioral signatures* as the primary scientific object replacing single-number accuracy, and *per-action ablations* that decompose a controller by representing it as a small explicit action set. They produce within-controller and within-component contrasts, most robustly majority lock-in, that reproduce across the matrix even where accuracy rankings

do not. A per-action ablation is diagnostic only where a component fires often enough to measure; where one is dormant, the methodology reveals that fact rather than producing a misleading null.

- Operationalization of three classical leadership theories as process-level controllers: the first substantive case study of the methodology, with team science supplying the principled substrate for component-decomposable action sets.
- A decomposition of each controller's gain into *recovery* and *breakage*: recovery repairs incorrect round-0 majorities and breakage corrupts correct ones, making the internal dynamics of multi-agent deliberation legible where a single accuracy number is silent. It reorganizes otherwise scattered controller-label results onto one measurable axis, the reliability of the independent round-0 majority, and isolates a *recovery advantage* over plain interaction as the precise condition under which a controller earns its structure, met in exactly one combination, the predicted apex of a characterized landscape rather than a lucky draw (Section 6).
- A mapping from the measured boundary back to leadership contingency theory (Section 6.3): a reliable round-0 acts as a *leadership substitute*, an unrecoverable task offers no latent ability to activate, undirected recovery makes control *redundant* (path-goal), and the single region where control helps is the *readiness gap* situational leadership predicts, so a largely null accuracy result is what the contingency view expects, not a failure of the controllers.

2 Related Work

Process- and knowledge-level coordination in multi-agent LLM systems. Recent multi-agent systems literature has explored several engineering patterns that improve over single-agent inference. *Multi-agent debate* (MAD) variants (Du et al., 2023; Liang et al., 2024; Chan et al., 2024) have agents independently propose, then iteratively critique and revise toward consensus, aggregating by majority vote or a final-round judgment. *Role specialization and pipelined decomposition* frameworks (Hong et al., 2024; Wu et al., 2024; Li et al., 2023; Chen et al., 2024) assign distinct personas, sub-tasks, or pipeline positions to different agents and route communication accordingly. *Self-refinement and critic-then-resolve loops* (Madaan et al., 2023; Shinn et al., 2023) use a single agent (or a critic head) to iteratively improve outputs through structured feedback. These mechanisms primarily operate at the *knowledge level*: they shape *what* each agent reasons about, *which* agent contributes which subtask, or *whether* a candidate answer is refined further before final aggregation; coordination is then implicit in the choice of debate format, decomposition rule, or aggregation procedure. Our work is complementary. We hold the agent set and the final aggregation step fixed and study a small *process-level* action vocabulary (when to explore, revise, accept, or synthesize, with policy-specific extensions) that shapes round-to-round interaction dynamics rather than the substantive content of agent reasoning. We compare against a peer-to-peer MAD baseline directly in Appendix B.8.

Behavioral measurement and component-level controller analysis. Most MAS evaluations report final accuracy, rounds, and token cost; trajectory-level behavioral metrics are reported irregularly and rarely treated as the primary measurement target (Du et al., 2023; Liang et al., 2024; Chan et al., 2024; Hong et al., 2024; Madaan et al., 2023). Process metrics that do appear (turn counts, agreement rates, retry frequencies) are typically descriptive add-ons rather than controller-discriminating instruments. Two methodological gaps follow. First, behavioral measurement is sharpest when paired with controllers whose action sets are explicit: signatures like lock-in and recovery contrast most cleanly when one knows which control action "locks in" or "reopens." Second, component-level ablations of MAS controllers are uncommon because controllers are typically specified as monolithic prompts; ablating a prompt fragment does not cleanly ablate a behavioral component. We address both gaps by representing controllers as theory-derived explicit action sets, which makes behavioral signatures comparable across controllers and per-action ablations clean by construction. That majority voting is a strong baseline for multi-agent LLM systems is increasingly recognized; our contribution is not that observation but the measurement that localizes it: a shared-round-0 attribution and a recovery/breakage decomposition that identify which controllers add value over voting and over plain interaction, where, and by what mechanism. A matched arbitrary controller (the same action set assigned at random) recovers no better than majority voting (Appendix B.9), confirming that the theory-derived rule, not the vocabulary, is what the per-action ablations isolate.

Team science as a motivating frame. A broader line of team-science and organizational research argues that team performance is not determined by structure alone (Hackman, 2002; Salas et al., 2005; Hilton & Cooke, 2015; Greer et al., 2018). Diversity can broaden coverage while increasing friction; hierarchy can improve accountability while also suppressing information flow (Cox & Blake, 1991; Horwitz & Horwitz, 2007; Van Knippenberg et al., 2004). These findings motivate a design language for agent teams that goes beyond "flat versus hierarchical" or "homogeneous versus diverse" and instead asks what mechanisms regulate conflict, revision, coordination cost, and adaptation once a team is already assembled (Muralidharan et al., 2025).

Leadership theories used as controller specifications. We choose three leadership theories because together they span three distinct axes of the operational leadership literature. Transactional leadership emphasizes standards, monitoring, and contingent feedback (Burns, 1978; Bass, 1985); transformational leadership emphasizes shared goal alignment coupled with structured exploration (Bass, 1985); situational leadership treats the appropriate style as a function of team state rather than a fixed identity (Hersey & Blanchard, 1969). The triad spans *focus* (correction vs. goal-framing), *level* (individual contingent feedback vs. collective objective), and *adaptivity* (fixed style vs. state-driven switching). Other styles (servant, authentic, inclusive) are useful at the value or relational level but do not introduce a new control axis at the operational level we model. We treat them either as extensions of the same control vocabulary (e.g., inclusive leadership as a dissent-preservation specialization, included as a probe in Appendix B.4) or as scope for future work.

Leadership is contingent, not universal. Team science has long held that leadership effectiveness depends on the situation, not the style alone. Contingency (Fiedler, 1967) and path-goal (House, 1971) theories tie a leader's value to situational favorableness and to whether the leader supplies something the team lacks; *functional* leadership (Hackman & Walton, 1986) frames the leader's job as doing "whatever is not being adequately handled for group needs"; and *leadership-substitutes* theory (Kerr & Jermier, 1978) identifies subordinate ability, experience, and intrinsically clear tasks as conditions under which leadership is unnecessary or neutralized. Situational leadership (Hersey & Blanchard, 1969) makes the dependence explicit, switching style with follower readiness. We adopt this contingency stance as our null hypothesis rather than assuming control helps: we expect process-level control to add value only in a restricted regime, and we use the measurement vocabulary above to identify which team states correspond to each predicted regime (Section 6.3).

3 Leadership as Coordination Control

This section specifies the process-level control framework that the methodology operates on, then instantiates it with three classical leadership styles as the first case study. Let a team of agents be denoted by $A = \{a_1, \ldots, a_n\}$. At interaction step t, each agent produces an intermediate output $y_{i,t}$ conditioned on task input x, interaction memory M_t, and the current team state S_t. We define a process-level control policy

$$\pi : (S_t, Y_t, M_t) \rightarrow U_t,$$

where $Y_t = \{y_{1,t}, \ldots, y_{n,t}\}$ and U_t is a set of control actions such as `revise`, `accept`, `explore`, or `synthesize`. A controller acts on the interaction process rather than only on prompt context, and it can be evaluated through its behavioral consequences. The leadership case study below uses three classical styles (transactional, transformational, situational) to instantiate the policy π with theory-grounded action sets; Section 3.3 adds the Bass two-component decomposition that the per-action ablations target.

3.1 Leadership as Control

All conditions share a common loop (Table 1). Agents first answer independently in round 0, then receive bounded memory plus the policy-conditioned control actions defined below. After each round, the team either synthesizes a final answer or continues until convergence or budget exhaustion. This keeps comparisons centered on controller behavior under a matched interaction scaffold rather than on changes in team structure, model choice, or execution budget.

The control space U_t contains a small set of interventions:

```
Controller Skeleton
Input:  task x, team outputs Y_0, state S_0
for round t = 0,...,T:
  assess disagreement / rationale conflict / early consensus
  if policy = transactional:  revise or accept, then synthesize
  if policy = transformational:  explore via differentiated roles, then synthesize
  if policy = situational:  choose explore vs.  converge from team state
  update memory M_t and team state S_t
return final team answer
```

Table 1: Compact view of leadership as control. The policies differ not in agent identity or prompt role alone, but in which control actions they trigger and when.

- `explore`: instruct an agent to pursue a distinct line of reasoning, evidence source, or decomposition strategy;
- `revise`: require an agent to update its current answer based on criticism or mismatch with the team objective;
- `accept`: mark an intermediate answer as good enough to enter the synthesis pool without further revision;
- `synthesize`: aggregate accepted intermediate outputs into a team-level answer;
- `justify`: request explicit explanation or evidence from an agent before the output can be accepted.

Policies differ in how often they invoke these actions, which is itself part of the evaluation. Appendix C gives compact pseudocode for each policy so that the operational differences are easy to inspect.

3.2 Policy Instantiations and Baselines

Transactional leadership. The transactional controller implements explicit oversight and criterion-based feedback (Burns, 1978; Bass, 1985). For each agent:

$$u_{i,t} = C(y_{i,t}, r_t),$$

where C is a critic function and r_t is the current acceptance criterion. If $u_{i,t} = 0$, the agent revises; if $u_{i,t} = 1$, the output is accepted into the synthesis pool. In practice this produces a centralized revise–accept loop with strong convergence pressure.

Transformational leadership. The transformational controller implements shared goal framing plus structured exploration (Bass, 1985). The leader broadcasts a shared *goal* g_t (the *broadcast goal*) while assigning differentiated exploration directives $d_{i,t}$:

$$x'_{i,t} = f(x, g_t, d_{i,t}, M_t).$$

The purpose is coordinated variation around the broadcast goal rather than unconstrained divergence. Operationally, the leader intervenes less through local correction and more through framing, complementary exploration roles, and delayed synthesis.

Situational leadership. Situational leadership is an adaptive controller (Hersey & Blanchard, 1969) that switches between exploration-oriented and convergence-oriented behavior based on team state rather than task identity. The switch uses fixed, low-degree-of-freedom thresholds rather than tuned parameters. Concretely, for a 3-agent team the controller opens one exploratory step before synthesis whenever the round's answers are not unanimous (a three-way split, or a 2–1 split) or are nominally unanimous but show genuine underlying disagreement, operationalized as at least three distinct rationale openings (first-sentence prefixes) together with conflict markers (e.g., "however," "contradict," "unlikely") from at least two agents. Otherwise it falls back to a transactional-style convergence step, and it closes the round once a 2-of-3 majority agrees (Appendix C gives the full rule). This keeps the policy interpretable: its goal is not to invent a new specialist strategy, but to decide when the team should reopen deliberation and when it should close it.

Baselines. We compare these policies against two baselines.

- **Flat**: symmetric interaction and final aggregation by vote or simple synthesis.
- **Theory-free control**: a leader performs generic oversight using fixed critique loops or arbitration heuristics without theoretical grounding.

The theory-free control baseline is critical because it separates theory-grounded leadership from generic intervention. It uses a simple generic oversight rule: if the team has not yet converged, all agents are asked to revise; otherwise the team synthesizes. Unlike transactional or transformational control, it has no explicit theory of why disagreement should be reduced or preserved.

3.3 Bass Two-Component Decomposition

The transactional and transformational controllers each correspond to a two-component structure in Bass's leadership theory (Bass, 1985), which we make explicit because it renders both controllers *component-decomposable*. Transactional leadership pairs *contingent reward* (acknowledge and accept behavior that meets the criterion) with *management by exception* (intervene when behavior fails the criterion). In our control vocabulary, these map directly to the `accept` and `revise` actions, so we can ablate the management-by-exception action by holding `accept` active and removing `revise` from the control space. Transformational leadership pairs *inspirational motivation* (a shared, energizing collective objective) with *individualized consideration* (differentiated attention to each follower's contribution). In our vocabulary, these map respectively to the broadcast goal g_t and the per-agent directives $d_{i,t}$, and we ablate the broadcast-goal component by removing g_t while keeping $d_{i,t}$. Per-component ablations on both controllers are revisited empirically in later sections, where they isolate which component carries the behavioral signature in each regime and which becomes inactive, or actively harmful, as the task gets harder.

3.4 Implementation Assumptions

We focus on 3-agent teams so that traces remain comparable and interpretable. Within each benchmark–model run, all policy conditions share the same base model, configured generation settings, and task-type round budgets; Appendix B.5 lists the concrete settings used in the main and cross-model evaluations.

The policy rules were not tuned on a held-out set to maximize accuracy. Instead, we use simple, low-degree-of-freedom thresholds chosen to keep the controllers interpretable and faithful to their intended leadership styles. The thresholds are stated explicitly (the situational trigger above, and the full per-policy rules in Appendix C) so that the operational difference between controllers does not hide in prompt text. Several rules follow directly from team structure, such as 2-of-3 majority convergence in transactional control, while others are deliberately conservative and only trigger extra exploration under clear signs of unstable consensus, rationale conflict, or explicit objection. This is a proof-of-concept design choice: our goal is to test whether leadership theories can be operationalized as measurable coordination mechanisms before optimizing the resulting controllers for peak task performance. A sensitivity analysis (Appendix B.13) confirms that the situational controller's social-regime gain is robust to the conflict-marker lexicon and to the precise thresholds (it survives a lexicon swap and grows under threshold relaxation) and is localized to the 2–1-split reopening trigger rather than any tuned hyperparameter.

Operationalization, not unique mapping. The controllers above are operationalizations of their source theories rather than uniquely correct implementations. The released configs and code allow other groups to substitute alternative operationalizations (e.g., entropy-based exploration triggers, dissent-preserving inclusive variants) within the same control-action vocabulary. The scientific claim is not that any particular controller is the canonical embodiment of transactional or transformational leadership, but that this control-action vocabulary is rich enough to produce distinct, measurable behavioral signatures across model families.

3.5 Cross-Round Majority Extraction

Open-ended numeric tasks introduce an answer-extraction subproblem independent of controller logic: when an agent's final reasoning is long, a last-line single-pass extractor occasionally picks up a stray number from the trailing trace rather than the agent's actual `Final answer:` commitment. We use a cross-round majority extractor as the default finalize step in all open-ended numeric evaluations: it scans every agent-round `Final answer:` declaration, normalizes the extracted answers, and takes the majority across all rounds and agents, falling back to the single-pass extractor where no such declaration was produced. The fix is a *measurement* improvement decoupled from controller logic. All conditions including `flat` benefit from it, with the leadership controllers gaining the largest fraction of the recovered accuracy. Cross-controller comparisons on open-ended numeric tasks therefore remain on a common basis.

4 Experimental Scope

4.1 Task Regimes

We evaluate leadership control in three base shared-information regimes, plus a mixed workload that interleaves all three (Section 4.3), for four task regimes in total. In all cases, agents see the same task input and answer space but receive different reasoning profiles: short instruction prefixes that orient each of the three agents toward a distinct judgment strategy (e.g., counterexample-first, surface-match, plausibility-first; the per-regime prefixes are released with the configs). This isolates coordination under divergent judgment without introducing private evidence. We now describe the three base regimes, each posing a different coordination challenge.

Verification-heavy closed-ended QA. This regime is built from selected CommonsenseQA and StrategyQA items (Talmor et al., 2019; Geva et al., 2021). It measures whether explicit control improves reliability under tasks that reward careful checking and efficient convergence.

Abductive ambiguity. This regime uses AlphaNLI examples (Bhagavatula et al., 2019) in which multiple hypotheses remain initially plausible. It tests whether leadership can avoid premature convergence when early superficial plausibility is misleading.

Social-norm ambiguity. This regime uses broad Social Chemistry items (Forbes et al., 2020) drawn from `confessions`, `amitheasshole`, and `dearabby`, so that the evaluation reflects heterogeneous social judgments rather than one narrow norm domain. It tests whether leadership can manage that heterogeneity, especially when context-sensitive and clearly unacceptable judgments coexist.

4.2 Selection Protocol and Reporting Layers

Because the paper is regime-focused rather than leaderboard-focused, each benchmark family is reported through two layers: a *broad layer* used as an anti-cherrypicking evaluation set, and a narrower *curated layer* used when a cleaner mechanism readout is needed.

Broad layers are constructed by fixed rules rather than by policy outcome. Closed-ended uses a balanced CommonsenseQA+StrategyQA pool with only a small fixed exclusion list. AlphaNLI uses the stable reference pool directly at 100-scale. Social Chemistry keeps only `B/C` items with intermediate agreement from conflict-rich areas. Appendix A and Table 11 provide the regime-by-regime construction summary, exclusion logic, and source composition. In the main text, the important point is simpler: curated results sharpen mechanism visibility, while broad-layer results test whether the same pattern survives when the workload is widened.

4.3 Experimental Setup

The main comparison uses five conditions: (1) flat deliberation, (2) theory-free control, (3) transactional leadership, (4) transformational leadership, and (5) situational leadership; Section 7 adds two Bass component ablations (transactional accept-only, transformational directives-only).

We evaluate these conditions on three open-weight model families served through a single self-hosted inference backend: `gpt-oss-120b`, `llama-4-scout`, and `gemma-4-31B-it`.[1] Using one backend for every model avoids the cross-backend heterogeneity that a mix of locally-quantized and API-served models would introduce, so accuracy differences reflect model and controller behavior rather than serving conditions. The three families also span a capability range, which matters because the controllers' effect depends on how reliable each model's independent reasoning is. The main matrix is 3 model families × 4 task regimes (closed-ended QA, abductive ambiguity, social-norm ambiguity, and a mixed workload), 12 (model, regime) combinations; the same three models are the testbed for the MATH-500 Level 5 extension (Section 6.2).

Two evaluation tiers. We organize the evaluation into two tiers. The *main matrix* (the three regimes above plus the mixed workload, across the three models) covers regimes in which the independent round-0 majority is typically reliable; we use it to establish where control reduces to voting and to locate any accuracy gains. A second tier of *boundary probes* deliberately varies round-0 reliability and recoverability to characterize where control can help at all: a cross-domain extension to MATH-500 Level 5 and three further regimes (adversarial NLI, Winogrande, and a contested moral-judgment set), all introduced and analyzed together in Section 6. The probe tier is analyzed separately and is not pooled into the main-matrix results.

Shared round-0 baseline. All conditions share an identical round 0. We generate each agent's independent round-0 answer once per (task, seed) from a neutral, policy-agnostic prompt and reuse it across every condition; controllers act only from round 1 onward. The round-0 majority vote (R0-vote) is therefore the same controller-free starting point for every condition, so the difference between a controller's final answer and R0-vote measures what its round-1+ interaction (revision, reopening, synthesis) contributes rather than how round 0 was framed.

All experiments use the same episode/round execution loop starting from this shared round 0, after which the policy may issue control actions from the core control space of Section 3 (`explore`, `revise`, `accept`, `synthesize`, `justify`), with situational adding a `reconsider` extension; each policy uses a subset. The inclusive-leadership probe adds two further actions (`object`, `defend`) and is reported separately in Appendix B.4. We log task identifiers, seeds, agent outputs by round, control actions, final answers, and token counts. Each (model, regime) combination is evaluated over three seeds (closed-ended, AlphaNLI, and broad Social Chemistry use 100-item sets; the mixed workload aggregates 150 items), and the MATH-500 extension uses five seeds. All three models are decoded at `temperature`=0.35; the remaining generation settings (token limits, round budgets, top-p defaults) are detailed in Appendix B.5.

We also include a mixed-workload benchmark of 150 tasks: 50 closed-ended items, 50 AlphaNLI items, and 50 Social Chemistry items. Regime identity is not exposed to the controller, making this a direct test of adaptive leadership under heterogeneous deployment conditions.

4.4 Evaluation and Analysis

We evaluate each policy along three axes: task performance, coordination cost, and behavioral mechanism. Task performance is measured with exact-match accuracy. Coordination cost is measured through interaction rounds and token cost. Behavioral mechanism is measured from logged trajectories through exploration rate, dissent preservation, majority lock-in, and recovery from bad round-0 consensus. Dissent preservation (whether a correct minority answer survives into the final synthesis) is reflected in the recovery rate and probed directly by the inclusive-leadership ablation (Appendix B.4). The main tables foreground three highest-signal indicators, all per-run rates: *majority lock-in*, the fraction of items whose final answer equals the independent round-0 majority; *exploration*, the fraction of episodes in which the controller issues an `explore` action; and *recovery*, the fraction of incorrect round-0 majorities the controller repairs (P(final correct | round-0 majority wrong)).

[1] Names are the checkpoint identifiers exposed by that backend and correspond to public open-weight releases: `gpt-oss-120b` (OpenAI's open-weight 120B reasoning model), `llama-4-scout` (Meta Llama 4 Scout), and `gemma-4-31B-it` (a Google Gemma release in the ∼30B class). We report the gateway identifiers verbatim for reproducibility; exact revisions are pinned in the released configs.

Condition	gpt-oss-120b				llama-4-scout				gemma-4-31B-it			
	Acc.	Explore	Lock-In	Recovery	Acc.	Explore	Lock-In	Recovery	Acc.	Explore	Lock-In	Recovery
Flat	0.920	0.000	0.977	0.130	0.873	0.000	0.960	0.175	0.950	0.000	0.980	0.286
Theory-free control	0.903	0.000	0.963	0.087	0.880	0.000	0.953	0.225	0.920	0.000	0.990	0.000
Transactional leadership	0.923	0.000	**1.000**	0.000	0.867	0.000	**1.000**	0.000	0.930	0.000	**1.000**	0.000
accept-only ablation	0.923	0.000	1.000	0.000	0.867	0.000	1.000	0.000	0.930	0.000	1.000	0.000
Transformational leadership	0.903	0.100	0.963	0.087	0.873	0.153	0.953	0.200	0.927	0.063	0.990	0.048
directives-only ablation	0.917	0.100	0.973	0.087	0.880	0.153	0.960	0.200	0.920	0.063	0.990	0.000
Situational leadership	0.923	**0.167**	0.980	0.130	0.893	**0.340**	0.953	0.275	0.917	**0.273**	0.980	0.048

Table 2: Cross-model behavioral metrics on AlphaNLI ($n = 300$ per condition, 3 seeds). All three model families show the same ordering: transactional control has the highest majority lock-in (bold) with zero exploration, while situational shows the highest exploration (bold). The two indented rows are the Bass two-component ablations of Section 7.

4.5 Predictions

Because each controller operationalizes a leadership theory (Section 3), that theory's own claim about when the style is needed transfers to a prediction about the controller, fixed in advance rather than read off our implementation. Transactional leadership (Bass, 1985) enforces an existing standard but offers no mechanism to find a better one, so we predict it locks in the round-0 majority: efficient and matching the vote where that majority is reliable, brittle where it is wrong. Transformational (Bass, 1985) and situational (Hersey & Blanchard, 1969) leadership add, respectively, exploration around a shared goal and direction matched to the team's readiness to improve; both should break lock-in and recover incorrect majorities, but only where the majority is wrong, a better answer is reachable, and plain interaction has not already found it. Situational should gain most, since its readiness-based trigger reopens deliberation exactly when the team shows genuine disagreement. These predictions are inherited from theory rather than tuned to the data: together they imply a largely null accuracy result punctuated by a single localized gain, with the controllers separating on behavioral signatures (lock-in, exploration, recovery) even where accuracy does not. Section 6.3 returns after the results to ground each measured region in the contingency construct it instantiates.

5 Results

5.1 Behavioral Mechanism Analysis

Behavioral signatures are this work's primary scientific object, so we present them first. Table 2 summarizes the cross-model AlphaNLI readout for three high-signal indicators: explicit exploration, majority lock-in, and recovery from an incorrect round-0 consensus. Appendix A.1, Appendix A.2, and Appendix A.3 report the corresponding regime-level follow-ups.

The same ordering holds across all three model families, as Section 4.5 predicted: transactional control locks in the round-0 majority at or near 1.000 with essentially zero exploration and near-zero recovery, while the exploration-oriented controllers (situational, transformational) trade lock-in for higher exploration and recovery. Transactional's lock-in is the highest of any controller in every (model, regime) combination under a paired permutation test ($p < 0.05$; most $p < 0.001$). Figure 1 shows this across all (policy, model, regime) points: transactional and its accept-only ablation sit in the lower-right lock-in zone everywhere, while the other controllers spread toward higher recovery and lower lock-in.

The reopening profile is strongest on the weakest model, `llama-4-scout`, where situational reaches the highest exploration (0.34 on AlphaNLI, 0.42 on the social regime) and recovery (up to 0.29 on social). It compresses on the stronger `gpt-oss-120b` and `gemma-4-31B-it`, whose independent round-0 reasoning is reliable enough (AlphaNLI round-0 vote ≥ 0.92) that little incorrect consensus remains to recover from. The social regime is the clearest test of dissent preservation: on `llama-4-scout`, situational control reaches 0.513 accuracy[2] against transactional's 0.430 and flat's 0.433, with the lowest lock-in (0.727) and highest

[2]This headline combination appears as 0.513 in the main matrix, 0.517 in some appendix probes (Appendices B.8, B.4), and 0.527 in the budget-6 and sensitivity runs (Appendices B.11, B.13). The appendix probes are independent runs with their own

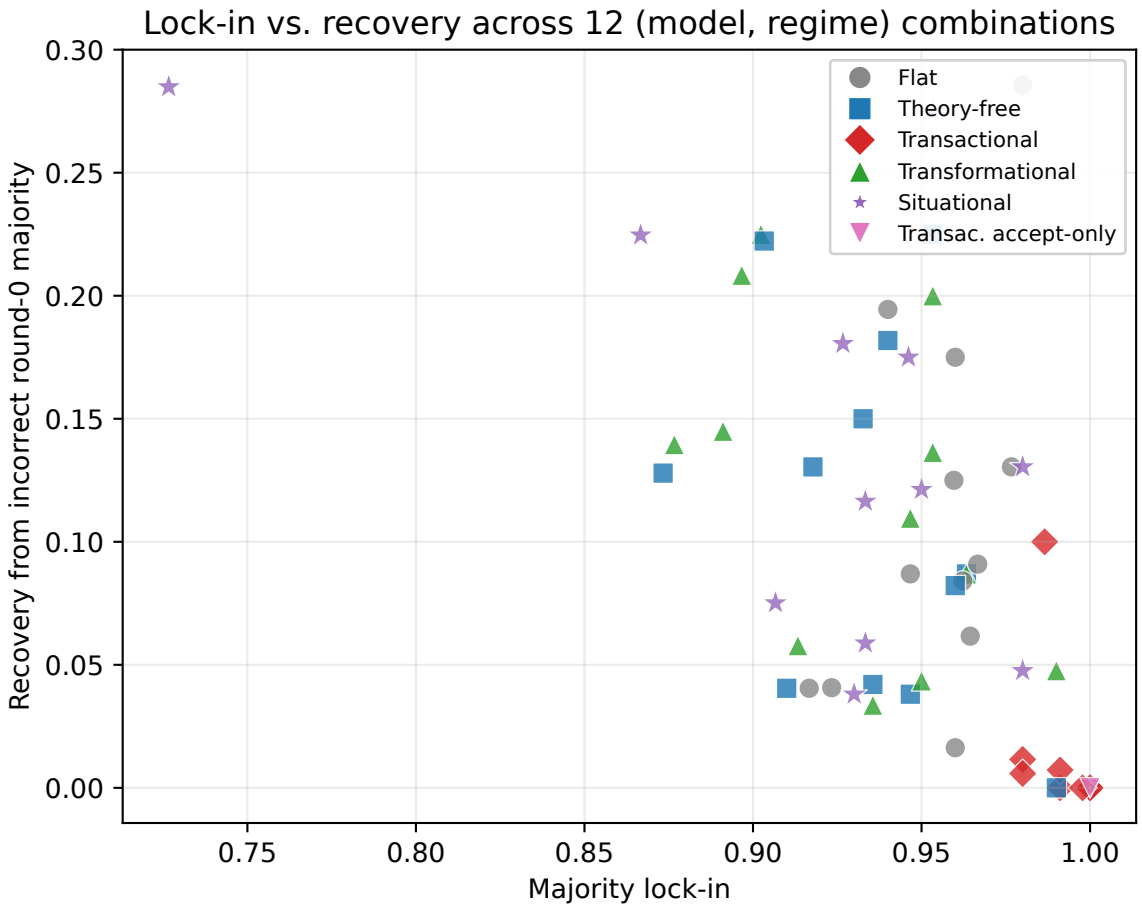

Figure 1: Lock-in vs. recovery across the 12 (model, regime) combinations (3 model families × 4 regimes). Transactional control and its accept-only ablation cluster in the lower-right lock-in zone in every combination; the other controllers spread toward higher recovery and lower lock-in.

recovery (0.285), and transformational reaches 0.483 (both gains significant over the shared round-0 vote after correction; Section 5.4). On the stronger models the social accuracies compress and no controller separates significantly, but the behavioral ordering persists.

5.2 Performance and Cost Comparison

The accuracy results complement the behavioral analysis. We read each controller against two baselines, weakest first. The *round-0 vote* is the simplest possible aggregation of the shared independent draws; beating it shows that a controller's round-1+ interaction adds anything at all. *Flat* is undirected multi-round deliberation; beating it is the stricter test, since it shows the controller's *structure* adds value over plain interaction. We report both, and treat the gain over flat as the decisive one. Table 3 reports the cross-model accuracy across the four regimes; Appendix B.2 reports stopping rounds, and Figure 2 the matched cost-quality Pareto view (mean tokens per run vs. accuracy for every (model, regime, condition) combination), with numeric token costs in Appendix B.3.

Table 3 shows why a single universal ranking is the wrong summary. The accuracy spread across conditions is small on every (model, regime) combination except `llama-4-scout` social, where situational (0.513)

generated-once shared round 0, so main-condition accuracies differ from the main matrix by at most ∼2pp; we keep each figure as-run rather than reconcile them.

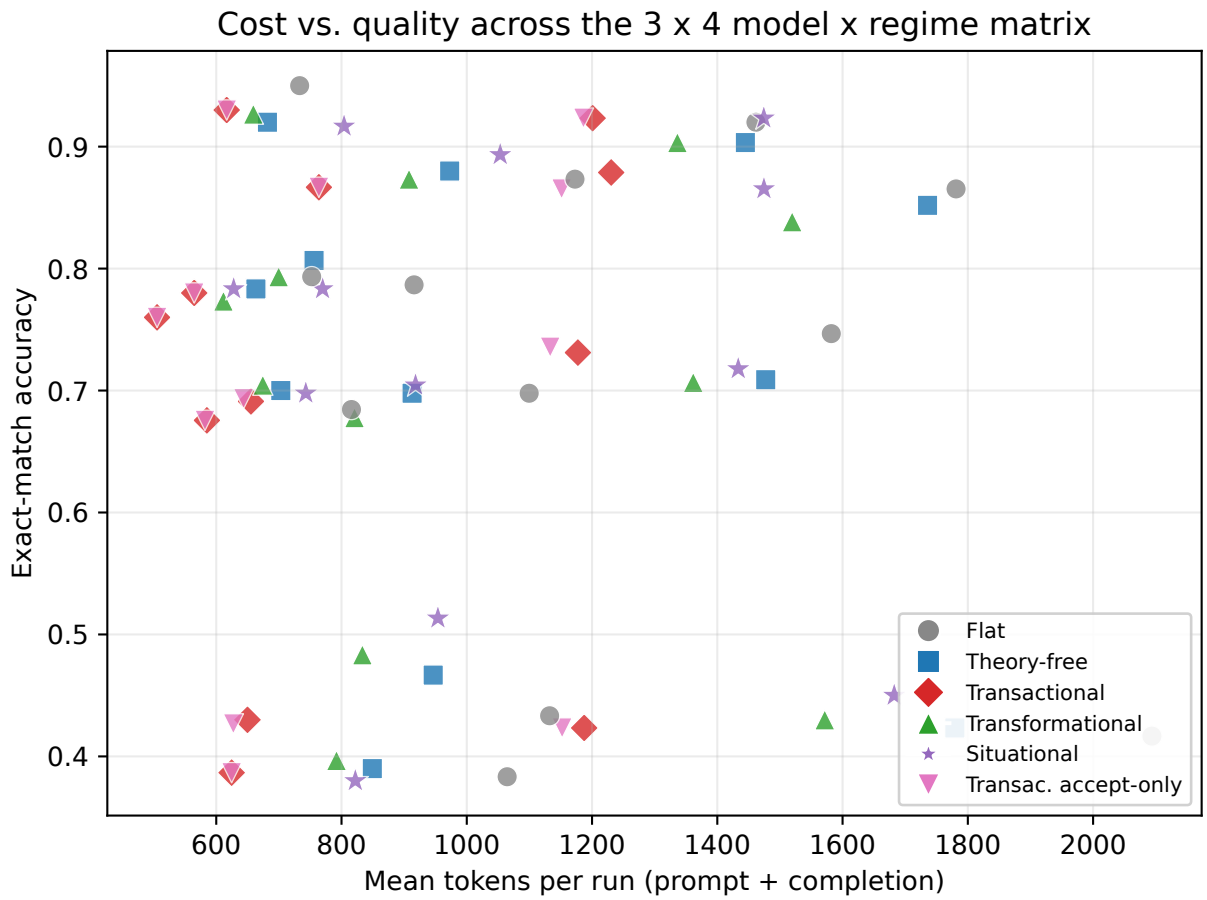

Figure 2: Cost-quality Pareto across the 3×4 model-family $\times$ regime matrix. Each marker is one (policy, model, regime) combination; x = mean prompt+completion tokens per run, y = exact-match accuracy. The transactional accept-only ablation terminates at round 0 by construction and is the cheapest condition; situational takes the Pareto frontier on the `llama-4-scout` social regime.

Condition	`gpt-oss-120b`				`llama-4-scout`				`gemma-4-31B-it`			
	Closed	AlphaNLI	Social	Mixed	Closed	AlphaNLI	Social	Mixed	Closed	AlphaNLI	Social	Mixed
Flat	0.865	0.920	0.417	0.747	0.787	0.873	0.433	0.698	0.793	0.950	0.383	0.684
Theory-free control	0.852	0.903	0.423	0.709	0.807	0.880	0.467	0.698	0.783	0.920	0.390	0.700
Transactional leadership	0.879	0.923	0.423	0.731	0.780	0.867	0.430	0.691	0.760	0.930	0.387	0.676
accept-only ablation	0.865	0.923	0.423	0.736	0.780	0.867	0.427	0.693	0.760	0.930	0.387	0.676
Transformational leadership	0.838	0.903	0.430	0.707	0.793	0.873	**0.483**	0.678	0.773	0.927	0.397	0.704
directives-only ablation	0.855	0.917	0.420	0.704	0.787	0.880	0.453	0.682	0.783	0.920	0.373	0.684
Situational leadership	0.865	0.923	0.450	0.718	0.783	0.893	**0.513**	0.704	0.783	0.917	0.380	0.698

Table 3: Cross-model exact-match accuracy across the four regimes ($n = 300$ for the 100-item regimes, $n = 450$ for the mixed workload; 3 seeds). Indented rows are the Bass two-component ablations of Section 7. Bold marks the only two leadership conditions whose gain *over the shared round-0 vote* is significant (`llama-4-scout` social; Table 4), not the within-column maximum, which is the usual convention elsewhere. Bootstrap 95% CIs and pairwise paired permutation p-values in Appendix B.6 and Table 15.

and transformational (0.483) separate from the rest. Transactional usually stops in the fewest rounds (Appendix B.2) but is never strongest in accuracy. On the easier regimes (closed-ended, AlphaNLI) the stronger models cluster near their round-0 ceiling and the controllers are statistically indistinguishable; on the mixed workload the nominal winner shifts by model (flat on `gpt-oss-120b`, situational on `llama-4-scout`,

transformational on `gemma-4-31B-it`), with differences inside paired-test noise. We therefore read the behavioral signatures, which are stable across the matrix, as the more reliable comparison than the absolute accuracy ordering. A peer-to-peer MAD baseline, run on all 12 combinations on the same shared round-0, never significantly beats a leadership controller: MAD tracks flat and the round-0 vote everywhere, and on the one combination with a real gain (`llama-4-scout` social) it loses to situational by 6.3pp ($p = 0.01$). A random-switch ablation further confirms that situational's social gain comes from its signal-driven trigger rather than its larger action space (Appendix B.8).

The cost picture (Figure 2) reinforces this reading. Accuracy is set mostly by regime and model rather than by controller, so the conditions differ mainly in tokens spent: the transactional accept-only ablation, which stops at round 0, is the cheapest condition and transactional control is close behind, while the exploration-oriented controllers (situational, transformational) spend more for their extra rounds. On the eleven combinations where the controllers stay within accuracy noise, that extra spend buys nothing, so the cheaper conditions sit on or near the cost-quality frontier; situational reaches the frontier only on `llama-4-scout` social, where its added cost is repaid by a real gain. The cost axis thus traces the same contingency boundary as the accuracy axis.

5.3 Mixed-Workload Robustness

The regime-separated results showed control helping in only one narrow regime, where the round-0 majority is unreliable, and reducing to voting elsewhere, with the controllers separating behaviorally rather than by accuracy. The next question is whether an adaptive controller can still recognize and help on that regime when the regimes are interleaved and their identity is hidden from it. To test that scenario, we construct a mixed 150-task benchmark that combines 50 closed-ended items, 50 AlphaNLI items, and 50 Social Chemistry items without exposing regime identity to the policy. The goal is to approximate a more realistic heterogeneous workload in which the controller must infer, from the task and the team's early interaction state alone, whether to converge quickly or preserve ambiguity.

The mixed-workload winner shifts by model and the aggregate gaps are small: flat leads on `gpt-oss-120b` (0.747), situational on `llama-4-scout` (0.704), and transformational on `gemma-4-31B-it` (0.704), each within paired-test noise of the runners-up. The behavioral picture, by contrast, is stable: transactional control keeps near-complete lock-in (0.99–1.00) and near-zero recovery on every model, while situational trades extra interventions for lower lock-in (0.87–0.93), higher recovery, and stronger performance on the social slice and the worst-source slice (Table 12). As on the separated regimes, the adaptive controllers help most where the round-0 majority is least reliable, and the accuracy ordering is less stable than the behavioral one.

5.4 Per-controller gains over the round-0 vote

The absolute accuracy ordering shifts by regime and model, but a sharper question is which combinations show a leadership controller adding accuracy over the simplest possible aggregation. We compare each controller's final accuracy to the shared round-0 majority vote (R0-vote; Section 4.3). Because round 0 is identical across conditions, a controller's gain over R0-vote measures exactly what its round-1+ interaction contributes over an independent draw plus a vote.

Table 4 reports the gains. Transactional control reduces to the round-0 vote on all 12 (model, regime) combinations: its final accuracy is within 1.3pp of R0-vote everywhere, so on the main matrix its lock-in is accept-driven and its `revise` action is dormant. Across the 36 leadership entries (three controllers $\times$ four regimes $\times$ three models), only two exceed a 5pp gain that survives Benjamini–Hochberg correction, both on `llama-4-scout` (the lowest-capability model) on the social regime, the one combination where its round-0 majority is unreliable (round-0 vote accuracy $R_0 \approx 0.43$): situational (+8.7pp; bootstrap 95% CI $[+3.3, +14.0]$; $q = 0.05$) and transformational (+5.7pp; CI $[+2.3, +9.3]$; $q = 0.05$). The stronger `gpt-oss-120b` and `gemma-4-31B-it` show no significant gain even on the social regime.

The behavioral signatures at the two gain combinations match the accuracy effect (Table 5) and map onto the controllers' theoretical action sets. On `llama-4-scout` social, situational control breaks majority lock-in (0.73 vs. transactional's 0.98), explores in 42% of episodes, and recovers from an incorrect round-0 majority 29% of

Model	Regime	R0-vote	Transac.	Transform.	Situa.
gpt-oss-120b	AlphaNLI	0.923	+0.000	−0.020	+0.000
gpt-oss-120b	Closed	0.865	+0.013	−0.027	+0.000
gpt-oss-120b	Social	0.423	+0.000	+0.007	+0.027
gpt-oss-120b	Mixed	0.736	−0.004	−0.029	−0.018
llama-4-scout	AlphaNLI	0.867	+0.000	+0.007	+0.027
llama-4-scout	Closed	0.780	+0.000	+0.013	+0.003
llama-4-scout	Social	0.427	+0.003	**+0.057***	**+0.087***
llama-4-scout	Mixed	0.693	−0.002	−0.016	+0.011
gemma-4-31B-it	AlphaNLI	0.930	+0.000	−0.003	−0.013
gemma-4-31B-it	Closed	0.760	+0.000	+0.013	+0.023
gemma-4-31B-it	Social	0.387	+0.000	+0.010	−0.007
gemma-4-31B-it	Mixed	0.676	+0.000	+0.029	+0.022

Table 4: Accuracy gain of each leadership controller over the shared round-0 vote on the 12 (model, regime) combinations. The R0-vote column is the shared round-0 majority accuracy. * marks a gain $\geq$ 5pp that survives Benjamini–Hochberg correction across all 36 leadership entries ($q < 0.05$). Transactional gains nowhere (within 1.3pp of R0-vote everywhere); both significant gains fall on the lowest-capability model on the social regime.

Condition (llama-4-scout social)	Lock-In	Explore	Recovery
Flat	0.923	0.000	0.041
Theory-free control	0.873	0.000	0.128
Transactional	0.980	0.000	0.006
Transformational	0.877	0.290	0.140
Situational	0.727	0.420	0.285

Table 5: Behavioral signatures on the llama-4-scout social regime, where the two controller-specific accuracy gains appear. Lock-In: P(final = round-0 majority). Explore: fraction of episodes with an explicit exploration action. Recovery: P(final correct | round-0 majority wrong).

the time, against transactional's near-zero recovery (0.01); transformational shows the same direction more mildly (lock-in 0.88, explore 0.29, recovery 0.14) via broadcast-goal exploration. Situational active reopening and transformational broadcast-goal direction are the levers that move accuracy beyond the round-0 vote; transactional, which only accepts, cannot.

The methodology therefore separates two dimensions the controller label collapses: each controller has a reproducible behavioral signature (transactional locks in, situational actively reopens, transformational preserves dissent under a broadcast goal), but an *accuracy* gain over the round-0 vote appears only where the round-0 majority is unreliable enough for reopening to pay off: here, the weakest model on the social regime ($R_0 \approx 0.43$). Where the round-0 majority is already reliable (closed-ended and AlphaNLI at $R_0 \geq 0.78$, and the stronger models throughout), every controller compresses to within a few points of R0-vote, and transactional coincides with it. That such gains are rare is what Section 4.5 predicted; Section 6 characterizes that boundary in detail, locating where a controller can beat plain interaction at all and why.

6 Characterizing the Boundary: Recovery Advantage and a Cross-Domain Extension

Section 5.4 found accuracy gains over the vote in only two of the 36 leadership entries, both on a single (model, regime) combination. This section characterizes that scarcity directly and sharpens it: against the stricter flat baseline, even that gain survives in only one combination. Using the boundary-probe tier (Section 4.3), we decompose a controller's gain into recovery and breakage, define the recovery advantage that decides when control beats plain interaction, test it across four probe regimes (including the cross-domain MATH-500 extension), and read the resulting boundary against contingency theory.

6.1 Why gains over plain interaction are rare: a recovery-advantage account

A controller's gain over the shared round-0 vote decomposes exactly into a benefit and a cost,

$$\text{gain over the vote} = P(r_0 \text{ wrong}) \cdot \text{recovery} \; - \; P(r_0 \text{ right}) \cdot \text{breakage},$$

where recovery $= P(\text{final correct} \mid \text{round-0 majority wrong})$ is the share of incorrect round-0 majorities the controller repairs, and breakage $= P(\text{final wrong} \mid \text{round-0 majority correct})$ is the share of correct majorities it drags to error. Each product weights a conditional rate by how often its case arises, so the two terms are the shares of all items the controller flips from wrong to right and from right to wrong; their difference is the controller's accuracy minus the vote's, which we report in percentage points. Pooled over the 1,194 incorrect-round-0 and 2,853 correct-round-0 items of the main matrix (Table 6, Figure 3), transactional control has recovery 0.007 and breakage 0.002: it essentially never moves, which is precisely why it coincides with the round-0 vote. Situational and transformational control do real work (recovery 0.140 and 0.106 at a 0.041 breakage cost), but a net gain requires the benefit to outweigh the cost, which for situational happens only once $P(r_0 \text{ wrong})$ exceeds about 0.23.

Condition	**Recovery**	**Breakage**
Flat	0.075	0.021
Theory-free control	0.097	0.034
Transactional	0.007	0.002
Transformational	0.106	0.041
Situational	0.140	0.041

Table 6: Recovery and breakage pooled over the 12 main-matrix combinations (1,194 incorrect-round-0 and 2,853 correct-round-0 items). Transactional neither recovers nor breaks (it is the round-0 vote); situational recovers the most.

This makes the two significant gain combinations a prediction rather than a coincidence: they are exactly the combinations where the round-0 majority is unreliable enough *and* the model can recover.

Beating the round-0 vote, however, is not the same as beating plain interaction. Flat deliberation also recovers some incorrect round-0 majorities, so the question that decides whether a *controller* earns its structure is sharper: does it repair incorrect majorities that flat leaves unrepaired? We therefore read the *recovery advantage* $\Delta_{\text{rec}} = \text{recovery(situational)} - \text{recovery(flat)}$. A controller's accuracy gain over flat appears only when Δ_{rec} is large and positive, which, as the boundary probes below show, is rare.

Boundary probes. To locate where an accuracy gain over flat is possible, we probe four further regimes chosen to vary the two governing quantities independently (full per-model results in Appendix B.12). Two have objective gold that keeps the 3-agent round-0 majority *reliable* even when individual agents are fooled: adversarial natural-language inference (ANLI-R3 (Nie et al., 2020)) and Winogrande (Sakaguchi et al., 2021). One has subjective, contested gold that makes the round-0 majority *unreliable but unrecoverable*: a moral-judgment regime of 99 balanced, genuinely contested `r/AmITheAsshole` verdicts from Scruples (Lourie et al., 2021) (median community agreement 0.56), where predicting the crowd's plurality is not derivable from the text by reasoning. The fourth is MATH-500 Level 5 (Section 6.2), whose round-0 is unreliable and *recoverable but already exploited by flat*. Table 7 lines these up against the one regime where a controller wins. The recovery advantage Δ_{rec} orders the accuracy gain across all of them: it is near zero wherever the round-0 majority is reliable (ANLI, Winogrande: little incorrect consensus to repair), near zero wherever the task is unrecoverable (Scruples, $\Delta_{\text{rec}} \leq 0.01$ on every model; `gemma-4-31B-it` social), and negative wherever flat already recovers (MATH, $\Delta_{\text{rec}} < 0$); it is large (+0.24) only on `llama-4-scout` social, the single combination where the model repairs incorrect majorities that flat cannot.

The recovery term depends on model capability, which is why the win is narrow even among unreliable-round-0 regimes. On the social regime `llama-4-scout` recovers 0.28 of its incorrect majorities against 0.18 breakage (net +8pp over the vote), whereas `gemma-4-31B-it` recovers only 0.04 against 0.08 breakage (net −1pp): it cannot repair its own social errors, so reopening does not pay off even though its round-0 is equally unreliable.

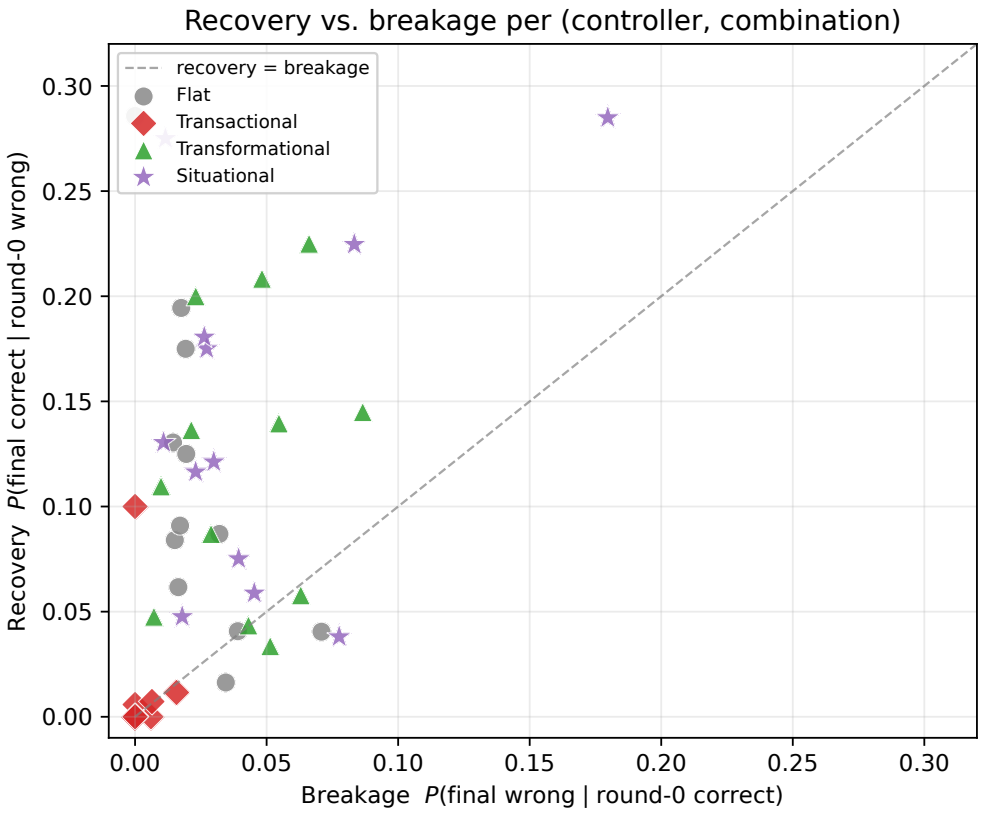

Figure 3: Recovery vs. breakage for each (controller, combination). Transactional control sits at the origin: it never moves, so it equals the round-0 vote. Situational and transformational lie above the $y = x$ line (they recover more than they break); the single situational point below the line is `gemma-4-31B-it` social, where the model cannot repair its own errors.

Regime	**Model**	$P(r_0$ wrong)	Recov. flat	Recov. situ.	Δ_{rec}	Δacc
Social	`llama-4-scout`	0.57	0.04	0.28	+0.24	+0.080*
Social	`gemma-4-31B-it`	0.61	0.02	0.04	+0.02	−0.003
Scruples	`llama-4-scout`	0.61	0.02	0.02	+0.00	−0.010
Scruples	`gpt-oss-120b`	0.55	0.03	0.04	+0.01	+0.000
ANLI-R3	`llama-4-scout`	0.27	0.21	0.22	+0.01	−0.063
Winogrande	`gemma-4-31B-it`	0.10	0.20	0.17	−0.03	−0.020
MATH-500	`llama-4-scout`	0.40	0.41	0.28	−0.13	−0.056
MATH-500	`gemma-4-31B-it`	0.22	0.76	0.39	−0.37	−0.088

Table 7: Boundary probes: a controller's accuracy gain over flat (Δacc = situational − flat, paired permutation) tracks its *recovery advantage* Δ_{rec} = recovery(situational)−recovery(flat), not round-0 unreliability $P(r_0$ wrong) alone. A leadership-specific gain requires all three of an unreliable round-0, a recoverable task, and a recovery flat does not already realize, met only on `llama-4-scout` social. Note that this Δacc is the gain over *flat* (+0.080 on that combination); the +8.7pp quoted elsewhere is the same combination's gain over the round-0 *vote* (Table 4), a weaker baseline. *: significant after correction.

Gains over plain interaction are therefore governed by quantities the controller label hides (how often the round-0 majority is wrong, whether the task is recoverable at all, and whether the controller recovers where

Condition	gpt-oss-120b				llama-4-scout				gemma-4-31B-it			
	Acc.	Lock-In	Recov.	Stop@R	Acc.	Lock-In	Recov.	Stop@R	Acc.	Lock-In	Recov.	Stop@R
Flat	0.960	0.947	0.625	1.24	0.736	0.632	0.414	1.76	0.944	0.792	0.759	1.58
Theory-free control	0.944	0.943	0.500	1.32	0.700	0.668	0.283	2.03	0.912	0.792	0.611	1.72
Transactional leadership	0.936	0.951	0.312	1.17	0.720	0.732	0.323	1.46	0.900	0.836	0.556	1.26
Transformational leadership	0.936	0.947	0.312	1.20	0.664	0.664	0.212	1.56	0.864	0.784	0.426	1.42
Situational leadership	0.936	0.955	0.312	1.31	0.680	0.648	0.283	1.87	0.856	0.776	0.389	1.64

Table 8: MATH-500 Level 5: accuracy and behavioral signatures, $n = 250$ per condition (5 seeds). Accuracy is full-sample with the cross-round majority extractor (Section 3.5); Lock-In and Recovery are computed over runs with a parseable round-0 majority; Stop@R is the mean stopping round. The `transactional_accept_only` ablation is reported in Table 10.

flat does not), and the methodology measures each of them. Their intersection is small, which is exactly why accuracy gains are rare; the recovery signature, not the controller label, is what locates it.

The recovery comes from the controllers' theory-derived decision rules. We verify this with an *arbitrary-control* baseline that uses the same control actions but assigns them uniformly at random each round, with no leadership-derived structure (Appendix B.9). On a matched run that adds it alongside the leadership controllers, the arbitrary controller recovers 0.009 of incorrect round-0 majorities: indistinguishable from the round-0 vote (0.010) and far below situational's 0.133 and transformational's 0.089 on the same items, with situational's recovery exceeding it in all 9 combinations. The same action set used arbitrarily has no effect; the theory-derived rule for *when* to reopen and *what* to broadcast is what produces recovery.

This also answers a natural objection: that situational control might be no more than a generic "reopen on disagreement" heuristic dressed in leadership terms. A signal-matched reopening rule is not a *non*-leadership control: state-contingent intervention *is* the operational content of situational leadership (Hersey & Blanchard, 1969). We bracket it from two sides. A matched-rate but signal-*blind* reopening rule (random-switch, Appendix B.8) underperforms situational, so the gain comes from reopening at the right times, not from reopening per se; and the sensitivity analysis (Appendix B.13) localizes the gain to the 2–1-split trigger. We make no claim that the differentiated directives beyond that reopening decision are themselves load-bearing; the theory's contribution is the decision rule for *when* to reopen.

The cross-domain extension below pushes round-0 reliability to its extremes, where this recovery advantage turns negative and pins down what a leadership-specific gain requires beyond an unreliable round-0.

6.2 Cross-domain extension: hard math (necessity is not sufficiency)

The 12 main-matrix combinations span regimes in which per-agent round-0 reasoning is typically reliable. We now apply the same behavioral-signature and per-action-ablation methodology to MATH-500 Level 5 (Hendrycks et al., 2021; Lightman et al., 2024), a competition-math benchmark (simple-numeric-answer subset, 50 items) on which round-0 reliability varies sharply across our three models. This is the regime in which the `revise` action of transactional control (Section 3.3), dormant on the main matrix, actually fires; Section 7 uses this to decompose transactional control per action.

6.2.1 Setup

We evaluate the three models with 3-agent teams under the five main conditions plus the `transactional_accept_only` ablation (Section 7), five seeds per condition ($n = 250$). The cross-round majority extractor from Section 3.5 is the default finalize step; `gpt-oss-120b`, a reasoning model, uses a larger generation budget so that its round-0 reasoning is not truncated before committing a final answer (Appendix B.5).

6.2.2 Behavioral signatures and accuracy on hard math

Table 8 reports per-condition accuracy and behavioral signatures across the three models. The three models span a wide range of round-0 reliability: the round-0 majority is incorrect in only ∼6% of `gpt-oss-120b`

runs, ∼22% on `gemma-4-31B-it`, and ∼40% on `llama-4-scout`. Where round-0 is reliable (`gpt-oss-120b`, round-0 vote 0.94), all controllers cluster near the flat ceiling (0.94–0.96) and deliberation adds little. Where it is unreliable (`llama-4-scout`, round-0 vote 0.60), multi-round deliberation recovers many incorrect round-0 majorities, lifting flat to 0.74 over the round-0 vote.

The behavioral ordering matches the main matrix: transactional control, and especially its accept-only ablation (Section 7), produces the highest lock-in and the lowest recovery, while flat and the exploration-oriented controllers recover more. Recovery itself scales with capability: every controller on `gpt-oss-120b` recovers from an incorrect round-0 majority more often than any controller on `llama-4-scout`. On `gpt-oss-120b` the leadership controllers nonetheless recover less than flat (0.31 vs. flat's 0.63): they close on the 2-of-3 majority and forgo the extra rounds flat spends repairing the rare incorrect one, but with only ∼6% of round-0 majorities wrong this costs almost no accuracy. The decisive comparison is the `revise`-action ablation, which isolates what the management-by-exception action contributes in each reliability regime (Section 7.2).

This MATH evidence sharpens the round-0-reliability axis into a *necessity-not-sufficiency* statement. Where round-0 is unreliable, interaction adds value, but on hard math that value accrues to *flat* interaction, not to the leadership controllers: flat is the most accurate condition on all three models, and every leadership controller trails it (situational by 2.4/5.6/8.8pp on `gpt-oss-120b`/`llama-4-scout`/`gemma-4-31B-it`; transactional, the lock-in controller, is the strongest or tied-strongest leadership condition on every model). This is the opposite of `llama-4-scout` social, where situational beats flat by 8pp (Table 3). The contrast pins down what a leadership-specific advantage requires beyond an unreliable round-0: a failure mode of premature consensus on genuinely ambiguous items, where structured reopening outperforms undirected interaction. On hard math the disagreement is largely self-correcting: a wrong numeric answer tends to expose its own error under scrutiny, so plain flat interaction already recovers incorrect round-0 majorities and leaves a controller little to add. Round-0 unreliability is therefore necessary but not sufficient for a leadership-specific gain; the gain also requires that undirected interaction cannot already exploit the unreliability on its own. We return to this boundary in the limitations (Section 8).

6.3 Mapping the boundary to leadership contingency theory

The boundary assembled across Sections 6.1–6.2 is not only an empirical regularity; it lines up with the contingency view of leadership we adopted as a null hypothesis. Team science holds that a leader adds value only by supplying a function the team needs and is not otherwise getting (Hackman & Walton, 1986), and that several conditions make leadership unnecessary or impossible (Kerr & Jermier, 1978). Each region of our measured boundary corresponds to one of those conditions (Table 9).

Measured region	**Signature**	**Contingency construct**	**Observe**
Round-0 reliable	$P(r_0$ wrong) low	leadership substitutes (Kerr & Jermier, 1978)	control = vote
Unreliable, unrecoverable	recovery ≈ 0	no latent ability to organize (Hackman & Walton, 1986)	control = vote
Unreliable, flat recovers	$\Delta_{\text{rec}} < 0$	path-goal redundancy (House, 1971)	control $\leq$ flat
Unreliable, recoverable, flat insufficient	$\Delta_{\text{rec}} \gg 0$	readiness gap (Hersey & Blanchard, 1969)	situational $>$ flat (+8pp)

Table 9: The measured boundary (Section 6.1, Table 7) read against leadership contingency theory. Each region instantiates a named construct, and the observed outcome confirms the predictions of Section 4.5; a controller adds value only in the readiness-gap region.

A reliable round-0 is the analogue of *leadership substitutes*: the agents already solve the task, so control reduces to the vote; an unrecoverable task (Scruples; `gemma-4-31B-it` social) offers no latent ability for a controller to organize; and where undirected interaction already recovers (hard math), directive control is *redundant* in the path-goal sense. A controller earns its keep only in the remaining *readiness gap* (an

unreliable round-0 that the team can repair but flat does not), the one region (`llama-4-scout` social) where situational control gains.

All four regions of this boundary follow from contingency theory and match the predictions of Section 4.5. The reliability conditions motivated the shared-round-0 design, and the boundary probes bear out the recoverability conditions, contrasting the unrecoverable Scruples probe with recoverable math. We offer the mapping modestly: the measurement vocabulary is a *computational testbed* whose boundary is consistent with leadership contingency theory, not an adjudication between theories.

7 Bass Two-Component Empirical Decomposition

Section 3.3 introduced two component-level ablations: `transactional_accept_only`, which removes the management-by-exception action from transactional control while keeping the contingent-reward action, and `transformational_directives_only`, which removes the broadcast-goal component from transformational control while keeping the differentiated-directive component. Each ablation isolates one component of a Bass classical leadership style at the action level, leaving the other component active. We report the transactional decomposition across the 12 main-matrix combinations and on MATH-500 Level 5 (Section 6.2), together with the transformational two-component ablation across the four regimes as a design-symmetric companion.

7.1 Main matrix: the management-by-exception action is structurally dormant

On the main matrix the `revise` (management-by-exception) action is structurally dormant. Transactional control stops at round 0 in nearly every episode (its mean stopping round is at most 1.03 on all 12 combinations), so the `revise` action is almost never available to alter the outcome. Two consequences follow. First, the transactional lock-in signature is carried entirely by the `accept` (contingent-reward) action: accept-only and full transactional converge in the same round 0 and accept-only inherits the same near-ceiling lock-in (Table 2). The signature is therefore *accept-driven, not revise-driven*: a strictly stronger claim than "revise contributes little," since revise is dormant by construction.

Second, with revise dormant the accept-only ablation coincides with the shared round-0 vote on all 12 combinations: the accept-only rows of Table 3 match the R0-vote column of Table 4 exactly. This confirms the controller-level reducibility result of Section 5.4 at the component level: where its second action never fires, transactional control adds nothing over voting the shared round 0.

7.2 Hard math: the revise action helps where round-0 is unreliable

On MATH-500, where round-0 disagreement is common, the `revise` action fires and its effect becomes measurable. Table 10 contrasts full transactional with its accept-only ablation across the three models. Removing the revise action lowers accuracy on the two models whose round-0 majority is unreliable, `llama-4-scout` (-8.0pp, $p < 0.01$) and `gemma-4-31B-it` (-8.0pp, $p < 0.01$), because the revise step is what recovers incorrect round-0 majorities through later rounds. On `gpt-oss-120b`, whose round-0 majority is already reliable (round-0 vote 0.94), removing revise has no effect: accept-only (0.928), full transactional (0.936), and the round-0 vote (0.935) coincide, all just below flat (0.960).

Model	R0-vote	Flat	Transac.	Accept-only	Δ_{revise}	p
`gpt-oss-120b`	0.935	0.960	0.936	0.928	-0.008	0.68
`llama-4-scout`	0.604	0.736	0.720	0.640	-0.080	<0.01
`gemma-4-31B-it`	0.784	0.944	0.900	0.820	-0.080	<0.01

Table 10: MATH-500 Level 5: effect of removing the `revise` action, $n = 250$ per condition. Δ_{revise} = accept-only − full transactional (paired permutation). Removing revise hurts the two models with unreliable round-0 majorities and has no effect on `gpt-oss-120b`, whose round-0 is already reliable; it never helps.

The revise action thus helps exactly where the round-0 majority is unreliable and has no effect where it is reliable: the same round-0-reliability threshold that makes the controller reducible to the round-0 vote on the

main matrix. Crucially, removing revise never improves accuracy at any model scale: there is no direction reversal in which convergence pressure becomes a net asset on a stronger model. Where round-0 is reliable, the controller simply has nothing to add; where it is unreliable, the revise action does real work.

7.3 Complementary decomposition: transformational two-component

The same design applies to the transformational controller, decomposed into the broadcast-goal preamble and the differentiated-directive per-agent loop (Section 3.3); the directives-only ablation rows appear in Table 3. The broadcast-goal component contributes on the ambiguity-heavy regimes and is null elsewhere: full transformational exceeds directives-only by +3.0pp on `llama-4-scout` social, +2.4pp on `gemma-4-31B-it` social, and +2.0pp on `gemma-4-31B-it` mixed, while on closed-ended QA and AlphaNLI the two coincide within ±2pp.

The two decompositions are regime-dependent at different mechanistic layers. The transactional decomposition is regime-dependent in *whether the second action fires at all* (revise dormant on the main matrix, active on unreliable-round-0 math); the transformational decomposition is regime-dependent in *which component carries the regime* (the broadcast goal on the ambiguity-heavy regimes, the differentiated directives elsewhere). Both land in the same Bass two-component frame.

7.4 Synthesis

Combining the decompositions, the management-by-exception (`revise`) action is dormant where the round-0 majority is reliable (the main matrix, and `gpt-oss-120b` on hard math) and decisive where it is unreliable (`llama-4-scout` and `gemma-4-31B-it` on hard math), tracking the round-0-reliability axis of Section 5.4. The transformational broadcast-goal component shows the same regime dependence, contributing on the social and mixed regimes and null elsewhere.

What a per-action ablation reveals therefore depends on whether the component fires. Where it does, the methodology yields a mechanism-level finding (the revise action's reliability-gated value); where it does not, it yields a reducibility result (transactional equals the round-0 vote) that bounds how strongly the controller-level signature can be claimed. The two outcomes are the same measurement applied in opposite reliability regimes.

8 Discussion

Contingency, confirmed and measured. Our results support the contingency view we began with: there is no universal best controller, and process-level control adds value only in the restricted readiness-gap regime that the measurement vocabulary identifies and team-science theory predicts (Sections 6.1–6.3). Across the three model families, absolute accuracy rankings shift more than the within-controller behavioral signatures, so the stable scientific object is the interaction signature a controller induces, not a single best policy.

Three patterns of regime fit. *Transactional control* is the lock-in mechanism: its majority lock-in is the highest of any controller in every (model, regime) combination ($p < 0.05$, most $p < 0.001$; Appendix B.6), which makes it efficient for disciplined convergence and brittle under genuine ambiguity: two faces of the same accept-driven mechanism, and the reason it coincides with the round-0 vote wherever that vote is already good. *Situational control* is the clearest ambiguity-oriented mechanism: it produces the study's only significant accuracy gain (+8.7pp over the shared round-0 vote on `llama-4-scout` social, $q = 0.05$) by breaking lock-in and recovering from incorrect round-0 majorities. *Transformational control* shows the same direction more mildly (+5.7pp on the same combination) via broadcast-goal exploration. On the stronger `gpt-oss-120b` and `gemma-4-31B-it`, whose round-0 majorities are already reliable, all three controllers converge toward the round-0 vote and separate only in their behavioral signatures. In deployment terms, process-level control earns its token overhead only where the independent round-0 majority is unreliable yet recoverable; elsewhere plain voting or flat interaction already sits on the cost-quality frontier.

Component axis: Bass two-component decomposition is regime-dependent. The third axis is internal to each controller. The Bass two-component ablations show that the transactional controller's main-matrix signature is carried entirely by its *accept* (contingent-reward) action: the `revise` action is dormant (transactional stops at round 0 in nearly every episode), so removing it leaves both the lock-in signature and the round-0-vote accuracy intact. On MATH-500 Level 5, where round-0 majorities are unreliable for the weaker models, the `revise` action does fire and removing it lowers accuracy by $\sim$ 8pp ($p < 0.01$); on the model whose round-0 is already reliable it has no effect, and it never helps. Transformational control shows the complementary pattern: its broadcast-goal component contributes on the ambiguity-heavy regimes and is null elsewhere. The component axis therefore obeys the same round-0-reliability rule as the controller axis: a second action earns its keep only where the first leaves an unreliable majority to repair.

Process-level scope, knowledge-level complementarity. The current study is intentionally restricted to *process-level* coordination on symmetric-input shared-information teams of three agents. *Knowledge-level* coordination patterns from the MAS literature (multi-agent debate, role specialization, pipelined decomposition, self-refinement (Du et al., 2023; Liang et al., 2024; Chan et al., 2024; Hong et al., 2024; Wu et al., 2024; Li et al., 2023; Chen et al., 2024; Madaan et al., 2023; Shinn et al., 2023)) are deliberately not the manipulated variable here. They remain complementary and can compose with the process-level vocabulary we develop. Asymmetric topologies (subtask decomposition, sequential pipelines, specialist-generalist splits) and agentic tasks that require tool use or multi-step execution introduce additional design degrees of freedom that would each plausibly explain any observed signature difference. Characterizing controller mechanism in the cleanest possible setting is a prerequisite for interpreting controller effects in those more confounded deployments.

Leadership as a starting point. Leadership theory is one entry into a broader research program that imports theory from team science into multi-agent LLM coordination. Conflict regulation, expertise recognition, trust calibration, and adaptive teaming each carry their own theoretical structure that could be mapped onto explicit control actions and probed with the same behavioral-signature and per-action ablation vocabulary developed here. Whether each admits such a faithful mapping in LLM agents, rather than only in human teams, is itself an empirical question that each case study must establish, not assume. We see leadership as a starting point rather than a destination.

Limitations. The paper has several limitations. The evaluation is built around regime-focused subsets and reference pools rather than full benchmark averages. This fits the mechanism goal, and we provide curation logs, bootstrap CIs, and paired permutation tests with multiplicity correction to limit overclaiming. Where accuracy gaps lie within the per-combination paired-test noise floor (Appendix B.6), we report mechanism contrasts rather than accuracy wins.

The positive accuracy signal concentrates on a single (model, regime) combination (`llama-4-scout` social). Section 6 characterizes this as the predicted apex of the boundary rather than a lucky draw, so we rest the broader claim on the recovery account (Section 6.1) and the arbitrary-control baseline (Appendix B.9), which hold across the matrix and on hard math, rather than on a wide set of accuracy wins.

Several scope limits remain. All three models are served through a single open-weight backend, which removes cross-backend serving confounds but leaves coverage of proprietary and still-larger models to future work. The main study fixes the team size at three and the round budget at 3–4 rounds; Appendix B.10 and Appendix B.11 report 5-agent and budget-6 ablations. Situational control currently relies on count- and rationale-divergence signals rather than entropy- or self-consistency-based uncertainty estimation; richer signal designs are future work. The paper focuses on short-horizon collective judgment, leaving long-term trust calibration, repeated episodes, and human evaluation of perceived leadership quality for future work.

9 Conclusion

The contribution is to treat process-level coordination control as a *contingency* question answered by measurement: not whether leadership controllers win, but under what conditions they add value, and whether those conditions match what team science predicts. Behavioral signatures (lock-in, exploration,

recovery) replace single-number accuracy as the primary scientific object, and per-action ablations expose which controller components carry the effect in which regime. The measured boundary maps cleanly onto contingency theory: leadership substitutes, path-goal redundancy, and the situational readiness gap (Section 6.3). A largely null accuracy result is the predicted outcome, not a failure of the controllers. The methodology is theory-grounded but theory-independent: leadership controllers are a first substantive case study, and the same measurement vocabulary applies to other team-science concepts (transactive memory, conflict regulation, trust calibration) treated as future case studies.

On the leadership case study, the methodology reveals structure beneath the controller labels: against a shared round-0 vote, transactional control reduces to the vote on all 12 main-matrix combinations, and the only significant accuracy gains fall on the one combination where an unreliable round-0 is also recoverable and not already repaired by plain interaction. A controller's interaction earns its keep precisely there: the readiness gap leadership contingency theory predicts, and the regime our measurement vocabulary is built to detect.

Ethics Statement

This paper studies coordination policies for multi-agent LLM teams on reasoning and social-judgment tasks. The main ethical risk is overinterpreting social or organizational metaphors as if they guaranteed desirable behavior. Our use of leadership terminology is operational rather than normative: we model leadership as explicit coordination control so that its behavioral consequences can be inspected, measured, and compared. A related risk is anthropomorphism: this language can encourage readers to over-attribute human-like intentions, social understanding, or organizational competence to the agents studied here. We do not claim that human leadership theories transfer wholesale to agent systems or that any single controller is universally beneficial.

The evaluation also includes social-norm and moral-judgment style data. Such datasets can reflect annotation artifacts, culturally narrow assumptions, or unstable judgments about acceptable behavior. To reduce that risk, we separate broad layers from narrower curated subsets, document fixed inclusion and exclusion rules, and report when effects weaken on broad layers.

The systems studied here are not intended for autonomous deployment in high-stakes settings. They are lightweight research controllers evaluated in short-horizon simulated team interactions. Any use of similar methods in consequential domains would require domain-specific validation, stronger uncertainty estimation, and substantially more careful human oversight than provided in the present study.

Reproducibility Statement

We designed the paper to be reproducible at the level of data construction, execution protocol, and logged team behavior. All reported conditions share the same 3-agent execution loop, base prompt structure, stopping logic, and logging schema, with policy differences isolated to control decisions. We report regime-separated evaluations, broad layers, focused curated subsets, and a mixed benchmark so that readers can distinguish mechanism visibility from broader robustness. The same evaluation matrix is run across three open-weight model families (`gpt-oss-120b`, `llama-4-scout`, `gemma-4-31B-it`) through a single inference backend, allowing readers to separate controller-specific interaction signatures from model-family-specific ranking shifts. Round 0 is generated once per (task, seed) and reused across conditions, so the round-0-vote baseline is reproducible and identical across controllers.

The project includes scripts for dataset normalization, split construction, experiment execution, and behavioral-metric analysis. The paper also reports source composition, seed usage, and the key evaluation scales used in the final experiments. For curated subsets, we retain explicit construction rules and exclusion notes so that regime design remains auditable rather than implicit. The code, configs, and evaluation artifacts used for the reported experiments are provided as anonymized supplementary material; the archival DOI is withheld for double-blind review.

Two practical limits remain. First, all three model families are open-weight models served through one gateway; coverage of proprietary and larger models is left to future work. Second, some controllers, especially

the adaptive and dissent-preserving variants, remain lightweight policy implementations rather than optimized systems. We therefore present this paper as a reproducible mechanism study rather than a maximally tuned benchmark submission.

LLM Usage Disclosure

LLMs were used during the preparation of this work for limited writing and coding support, including editing prose, restructuring text, and helping implement experiment scripts. The key research ideas and scientific framing originated with the first author. LLMs were not used to generate final scientific claims, to replace author judgment in data curation, or to serve as evaluation judges for the reported results. All reported analyses, interpretations, and final manuscript content were verified by the authors.

References

Bernard M. Bass. *Leadership and Performance Beyond Expectations*. Free Press, Collier Macmillan, 1985. URL `https://openlibrary.org/books/OL2861879M/Leadership_and_performance_beyond_expectations`.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*, 2019.

James MacGregor Burns. *Leadership*. Harper & Row, 1978. URL `https://openlibrary.org/books/OL22822398M/Leadership`.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *International conference on learning representations*, volume 2024, pp. 9079–9093, 2024.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *International Conference on Learning Representations*, volume 2024, pp. 20094–20136, 2024.

Taylor H Cox and Stacy Blake. Managing cultural diversity: Implications for organizational competitiveness. *Academy of Management perspectives*, 5(3):45–56, 1991.

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.

Fred E. Fiedler. *A Theory of Leadership Effectiveness*. McGraw-Hill, 1967.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 653–670, 2020.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

Lindred L Greer, Bart A De Jong, Maartje E Schouten, and Jennifer E Dannals. Why and when hierarchy impacts team effectiveness: A meta-analytic integration. *Journal of applied psychology*, 103(6):591, 2018.

J Richard Hackman. *Leading teams: Setting the stage for great performances*. Harvard Business Press, 2002.

J Richard Hackman and Richard E Walton. Leading groups in organizations. In Paul S Goodman (ed.), *Designing effective work groups*, pp. 72–119. Jossey-Bass, 1986.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Paul Hersey and Kenneth H Blanchard. Life cycle theory of leadership. *Training & Development Journal*, 1969.

Margaret L Hilton and Nancy J Cooke. *Enhancing the effectiveness of team science*. National Academies Press, 2015.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pp. 23247–23275, 2024.

Sujin K Horwitz and Irwin B Horwitz. The effects of team diversity on team outcomes: A meta-analytic review of team demography. *Journal of management*, 33(6):987–1015, 2007.

Robert J House. A path goal theory of leader effectiveness. *Administrative science quarterly*, pp. 321–339, 1971.

Steven Kerr and John M Jermier. Substitutes for leadership: Their meaning and measurement. *Organizational behavior and human performance*, 22(3):375–403, 1978.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems*, 36:51991–52008, 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.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 13470–13479, 2021.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Rasika Muralidharan, Haewoon Kwak, and Jisun An. Can lessons from human teams be applied to multi-agent systems? the role of structure, diversity, and interaction dynamics. *arXiv preprint arXiv:2510.07488*, 2025.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4885–4901, 2020.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Eduardo Salas, Dana E Sims, and C Shawn Burke. Is there a "big five" in teamwork? *Small group research*, 36(5):555–599, 2005.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019.

Regime	Broad Layer	Construction Rule	Main Broad Scale	Curated Logic
Closed-ended shared-information QA	CommonsenseQA + StrategyQA reference pools	Preserve defended anchor items, extend with stable reference-pool items, balance sources 50/50, remove a small fixed list of benchmark-fragile or low-diagnostic items	100 items	Smaller curated subsets retain only items with meaningful disagreement pressure and defensible gold labels
Abductive ambiguity	AlphaNLI stable reference pool	Start from the stable 100-item reference pool and apply no additional item-level filtering at 100-scale	100 items	Curated subsets retain items where both hypotheses are plausible and causal ambiguity is preserved
Social-norm ambiguity	Social Chemistry broad `B/C` pool	Keep only `B/C` ternary labels, dev split, action-agree in [2.0, 3.0], and areas `confessions`, `amitheasshole`, `dearabby`; prioritize lower-agreement items and preserve area diversity	100 items	Focused social cores retain items where context-dependent or unacceptable judgments remain interpretable without annotation artifacts

Table 11: Summary of the regime-construction protocol. Each broad layer is paired with a smaller curated set used to test whether the same qualitative policy pattern survives under tighter regime control.

Daan Van Knippenberg, Carsten KW De Dreu, and Astrid C Homan. Work group diversity and group performance: an integrative model and research agenda. *Journal of applied psychology*, 89(6):1008, 2004.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*, 2024.

A Curation Protocol and Reporting Layers

This appendix documents the evaluation construction logic used throughout the paper. Our goal is not to optimize subsets for any single policy, but to isolate coordination regimes that are otherwise diluted by trivial items, annotation-sensitive items, or task formulations that do not induce the intended interaction pressure. For each regime, we therefore distinguish between a broad layer and a narrower curated layer used for more mechanism-focused analysis.

A.1 Closed-Ended Shared-Information QA

The broad closed-ended regime is constructed from balanced CommonsenseQA and StrategyQA reference pools. The 100-item set preserves six defended anchor items from early pilot runs and then extends them with stable reference-pool items while maintaining a 50/50 source balance. The only exclusions are a small fixed list of benchmark-fragile or low-diagnostic items identified before the final conference-scale runs. This broad set is intended to test whether the early challenge-set pattern survives once the regime is widened toward a more verification-heavy shared-information workload.

A.2 Abductive Ambiguity

The AlphaNLI broad layer is intentionally simpler. At 100-scale, we use the stable reference pool directly and do not apply additional item-level exclusions. This makes AlphaNLI the cleanest anti-cherrypicking layer in the paper. The smaller curated AlphaNLI cores remain useful for qualitative and mechanism-focused analysis because they preserve especially strong explanatory ambiguity, but the 100-item broad set is already broad enough to support the paper's main quantitative conclusions.

A.3 Social-Norm Ambiguity

The Social Chemistry broad layer uses a scalable ternary mapping and keeps only `B/C` judgments, where `B` denotes context-dependent social evaluation and `C` denotes unacceptable behavior. We exclude `A` items because they too often collapse into easy positive cases that do not require dissent handling. To reduce source-specific bias, the broad 100-item social set is balanced across three conflict-rich areas: `confessions`, `amitheasshole`, and `dearabby`. We further restrict to the dev split and to intermediate agreement levels (`action-agree` in [2.0, 3.0]) so that the social benchmark reflects genuine norm ambiguity rather than either trivial consensus or pure annotation noise.

A.4 Why We Report Both Broad and Curated Layers

The paper's main claim is about coordination mechanisms, not leaderboard optimization. That goal requires regimes in which the relevant coordination pressure is actually present. At the same time, curated regimes create a legitimate concern about post hoc selection. Reporting both layers addresses that concern directly. The broad layer answers the anti-cherrypicking question: does the qualitative pattern survive when we expand the regime? The curated layer answers the mechanism question: what happens when the regime is made maximally legible? We therefore treat broad and curated results as complementary rather than competing forms of evidence.

B Additional Robustness Results

B.1 Mixed-Benchmark Source Breakdown

The mixed benchmark combines 25 CommonsenseQA items, 25 StrategyQA items, 50 AlphaNLI items, and 50 Social Chemistry items without exposing source identity to the controller. Table 12 reports the source-wise accuracy for all three models.

The source-wise view explains the small aggregate gaps in Table 3: the controllers differ mainly on the social slice, the hardest source, where situational improves worst-source accuracy on `llama-4-scout` (0.513 vs. transactional's 0.427). On the stronger models the slices compress and the controllers track each other. Behaviorally, transactional control keeps the highest mixed-workload lock-in and the fewest rounds on every model (Table 13), while situational reopens more often, the same lock-in-vs-reopening tradeoff as on the separated regimes.

B.2 Stopping-Round Comparison for the Main Regime Runs

B.3 Token Cost

Table 14 reports mean prompt-plus-completion tokens per run, pooled over the four regimes. Transactional control and its accept-only ablation are cheapest because they stop at round 0; flat and situational spend the most, mirroring the stopping-round ordering (Table 13).

B.4 Inclusive Leadership on the Social Regime

We include inclusive leadership (a dissent-preservation specialization of the control vocabulary that explicitly requests and records minority objections before synthesis) as a probe on the broad social regime (3 seeds; an

Model	Condition	AlphaNLI	CSQA	SocialChem	StratQA	Worst
gpt-oss-120b	Flat	0.940	0.880	0.413	0.893	0.413
	Theory-free control	0.913	0.840	0.360	0.867	0.360
	Transactional leadership	0.940	0.880	0.393	0.840	0.393
	Transformational leadership	0.913	0.853	0.373	0.813	0.373
	Situational leadership	0.927	0.867	0.387	0.813	0.387
llama-4-scout	Flat	0.873	0.773	0.420	0.827	0.420
	Theory-free control	0.853	0.787	0.433	0.827	0.433
	Transactional leadership	0.860	0.760	0.427	0.813	0.427
	Transformational leadership	0.840	0.747	0.440	0.760	0.440
	Situational leadership	0.847	0.773	0.513	0.733	0.513
gemma-4-31B-it	Flat	0.913	0.787	0.353	0.787	0.353
	Theory-free control	0.880	0.853	0.393	0.800	0.393
	Transactional leadership	0.880	0.813	0.367	0.747	0.367
	Transformational leadership	0.893	0.867	0.400	0.773	0.400
	Situational leadership	0.887	0.840	0.413	0.747	0.413

Table 12: Source-wise accuracy on the mixed 150-task benchmark (3 seeds) for all three models. The controllers differ mainly on the social slice, the hardest source; situational improves worst-source accuracy most on the weakest model (llama-4-scout, 0.513 vs. transactional's 0.427), the same regime where it gains over the round-0 vote.

Model	Condition	Closed	AlphaNLI	Social	Mixed
gpt-oss-120b	Flat	1.40	1.16	1.57	1.26
	Theory-free control	1.43	1.17	1.50	1.26
	Transactional leadership	1.03	1.01	1.03	1.03
	Transformational leadership	1.26	1.10	1.34	1.17
	Situational leadership	1.21	1.19	1.39	1.21
llama-4-scout	Flat	1.25	1.24	1.40	1.33
	Theory-free control	1.29	1.23	1.48	1.38
	Transactional leadership	1.00	1.00	1.03	1.02
	Transformational leadership	1.18	1.15	1.29	1.24
	Situational leadership	1.30	1.34	1.45	1.37
gemma-4-31B-it	Flat	1.25	1.11	1.39	1.22
	Theory-free control	1.30	1.10	1.36	1.20
	Transactional leadership	1.00	1.00	1.00	1.00
	Transformational leadership	1.19	1.06	1.25	1.14
	Situational leadership	1.21	1.27	1.28	1.24

Table 13: Mean stopping round across the four regimes (3 seeds). Transactional control stops at round 0 in nearly every episode (Stop@R ≈ 1.0), while situational and transformational spend extra rounds reopening deliberation.

independent run with its own shared round 0, so main-condition accuracies differ from Table 3 by at most $\sim$2pp). Accuracy is 0.457 / 0.467 / 0.403 for inclusive on gpt-oss-120b / llama-4-scout / gemma-4-31B-it, against situational's 0.437 / 0.517 / 0.397 and flat's 0.413 / 0.420 / 0.383. Inclusive is competitive with situational on the stronger models, but on llama-4-scout, where the round-0 majority is least reliable, situational's active reopening (0.517) outperforms inclusive's dissent-preservation (0.467). Dissent preservation

Condition	gpt-oss-120b	llama-4-scout	gemma-4-31B-it
Flat	1713	1082	839
Transactional	1197	658	583
accept-only	1153	649	582
Transformational	1438	816	683
Situational	1507	923	749

Table 14: Mean tokens per run (prompt + completion), pooled over the four main regimes. Transactional and its accept-only ablation are cheapest (round-0 stop); flat and situational are the most expensive.

is a real mechanism, but realizing an accuracy gain still requires reopening the unreliable majority, not merely recording the minority view.

B.5 Execution Settings

The released configs make the execution settings explicit. All reported runs use 3-agent teams, the shared episode/round execution loop with a generated-once shared round 0 (Section 4.3), and task-type round budgets of 3 rounds for closed-ended tasks and 4 for integration-style tasks. All three models (`gpt-oss-120b`, `llama-4-scout`, `gemma-4-31B-it`) are served through the same self-hosted backend with `temperature`=0.35. Generation length is `max_tokens`=512 on the main matrix; the MATH-500 Level 5 evaluations use a 3-round budget for open-ended numeric tasks and `max_tokens`=1024, raised to 3072 for `gpt-oss-120b` (a reasoning model) so that its round-0 reasoning is not truncated before it commits a final answer. The experiment code sets no explicit top-p override or context-window cap, so provider defaults remain in force on those dimensions.

B.6 Statistical Tests Across All Combinations

For every (model, regime, condition) combination we report bootstrap 95% CIs over item-level scores (10^4 iters) and paired permutation tests over per-task scores (10^4 iters, pairing on common (task, seed)) for every pairwise condition comparison; the complete table is released as a supplementary file. Table 15 summarizes representative highlights, recomputed on the shared-round-0 runs.

Model / Regime	Pair (A vs. B)	Metric	Diff	p
`llama-4-scout` / social	situational vs. transactional	acc	+0.083	0.004
`llama-4-scout` / social	transformational vs. transactional	acc	+0.053	0.007
`llama-4-scout` / social	situational vs. flat	acc	+0.080	0.007
`gpt-oss-120b` / social	situational vs. transactional	acc	+0.027	0.113
`llama-4-scout` / mixed	transactional vs. situational	lock-in	+0.124	< 0.0001
`gpt-oss-120b` / mixed	transactional vs. situational	lock-in	+0.058	< 0.0001
`gemma-4-31B-it` / mixed	transactional vs. situational	lock-in	+0.064	< 0.0001

Table 15: Representative paired-permutation contrasts on the shared-round-0 runs (10^4 iters). Accuracy gains concentrate on `llama-4-scout` social (the unreliable-round-0 combination); lock-in differences between transactional and the other controllers are significant in every (model, regime) combination, confirming the behavioral signature is more stable than the accuracy ordering.

B.7 Reporting Philosophy

We include these appendix tables because the paper's main risk is not underpowered experimentation; it is overinterpretation of carefully designed regimes. Our solution is to make the regime design itself auditable. Broad layers, source-wise mixed breakdowns, process metrics, and paired statistical tests together provide a transparent trail from raw workload construction to the mechanism claims made in the main text.

B.8 MAD Baseline and Situational Random-Switch Ablation

We benchmark the leadership controllers against a stronger peer-to-peer baseline on the *full* main matrix, and separately isolate the signal-driven component of situational control. *MAD* is a peer-to-peer multi-agent debate baseline following Du et al. (2023): each round agents read peer answers and may revise, with no leader and no acceptance criterion; the final answer is a majority vote. We run MAD against `flat`, `situational`, and `transformational` on all 12 (model, regime) combinations on the same generated-once shared round 0 (3 seeds, Table 16); these are independent runs from the main matrix, so accuracies differ by at most ∼2pp.

Model	**Regime**	R0-vote	Flat	MAD	Situ.	Transf.
`gpt-oss-120b`	AlphaNLI	0.937	0.930	0.923	0.927	0.907
	Closed	0.863	0.870	0.877	0.857	0.857
	Social	0.413	0.410	0.410	0.427	0.430
	Mixed	0.713	0.724	0.722	0.716	0.718
`llama-4-scout`	AlphaNLI	0.870	0.870	0.873	0.897	0.883
	Closed	0.790	0.797	0.783	0.790	0.803
	Social	0.437	0.453	0.450	**0.513**	0.467
	Mixed	0.693	0.711	0.713	0.707	0.696
`gemma-4-31B-it`	AlphaNLI	0.927	0.937	0.930	0.920	0.917
	Closed	0.777	0.797	0.807	0.777	0.777
	Social	0.390	0.390	0.377	0.400	0.387
	Mixed	0.678	0.671	0.680	0.696	0.691

Table 16: Full-matrix MAD baseline (3 seeds, shared round 0). MAD never *significantly* beats a leadership controller; where it nominally edges them the margin is $\leq$ 3pp on reliable-round-0 regimes where no condition separates from the vote. On the one combination with a real gain, `llama-4-scout` social, MAD (0.450) tracks flat (0.453) and the vote (0.437) and loses to situational (0.513) by 6.3pp ($p = 0.01$).

The full-matrix comparison sharpens the contingency picture: MAD, the recognized peer-debate baseline, behaves like undirected `flat` interaction everywhere. It clusters at the round-0 ceiling on the reliable regimes (its $\leq$ 3pp nominal edges over a controller are never significant and fall where nothing separates), and crucially it *fails to capture the readiness-gap recovery* on `llama-4-scout` social, where the signal-driven situational controller beats it by 6.3pp. Generic peer debate is not a substitute for the signal-driven trigger.

Situational random-switch. To isolate that trigger, we replace situational's team-state signals with a fixed exploration probability $p = 0.45$ (matched to the empirical exploration rate) while keeping its control space, on AlphaNLI / social / mixed (3 seeds; an independent run, ∼2pp drift). On `llama-4-scout` social (the one unreliable-round-0 combination), signal-driven situational (0.517) beats random-switch (0.477) and flat (0.430), whereas on the reliable regimes the variants compress (`gpt-oss-120b`: 0.430 vs 0.417; `gemma-4-31B-it`: 0.393 vs 0.390). The advantage comes from *when* the controller reopens exploration, not from the larger action space, the same round-0-reliability pattern as the main matrix.

B.9 Arbitrary-Control Baseline

To test whether the recovery the leadership controllers produce comes from the theory-derived decision rule or merely from access to the control vocabulary, we add an *arbitrary-control* baseline: each round it assigns every agent a uniformly random action from {`accept`, `revise`, `explore`}, with no leadership-derived structure. It shares the action set of the leadership controllers but its decision rule is chance. We run it alongside flat, transactional, transformational, and situational control on the AlphaNLI, social, and mixed regimes across the three models (3 seeds, shared round 0). Table 17 reports recovery and breakage pooled over these nine combinations.

The arbitrary controller's recovery (0.009) is indistinguishable from transactional control (0.010), which is itself the round-0 vote: an explicit controller that uses the action vocabulary without a principled rule recovers nothing. It is in fact below flat (0.061), because random actions churn the team without coherently

Condition	Recovery	Breakage
Flat	0.061	0.029
Transactional	0.010	0.003
Transformational	0.089	0.037
Situational	0.133	0.043
Arbitrary control (random action)	0.009	0.003

Table 17: Recovery and breakage pooled over the AlphaNLI, social, and mixed regimes on the three models (1,021 incorrect-round-0 items). Despite the same action set, the arbitrary controller recovers no better than the round-0 vote (transactional); situational and transformational recover far more. Situational's recovery exceeds the arbitrary controller's in all 9 (model, regime) combinations.

reopening a wrong majority. The theory-derived controllers' recovery (situational 0.133, transformational 0.089) therefore reflects what the theory contributes, a decision rule for composing the action set, not the action set alone.

B.10 Team-Size Ablation

We extend the AlphaNLI and social regimes to $n = 5$ agents on all three models (3 seeds, $n = 300$ per condition) to test whether the controller signatures persist beyond the 3-agent main study (Table 18).

		AlphaNLI			Social		
Model	**Condition**	Acc	Lock-In	Explore	Acc	Lock-In	Explore
`gpt-oss-120b`	Flat	0.900	0.920	0.000	0.423	0.913	0.000
	Transactional	0.933	**1.000**	0.000	0.447	**1.000**	0.000
	Transformational	0.913	0.960	0.743	0.437	0.897	0.857
	Situational	0.933	1.000	0.090	0.420	0.960	0.170
`llama-4-scout`	Flat	0.883	0.947	0.000	0.440	0.883	0.000
	Transactional	0.863	**1.000**	0.000	0.433	**1.000**	0.000
	Transformational	0.893	0.943	0.400	**0.503**	0.817	0.513
	Situational	0.860	0.997	0.300	0.493	0.873	0.217
`gemma-4-31B-it`	Flat	0.953	0.940	0.000	0.360	0.923	0.000
	Transactional	0.920	**1.000**	0.000	0.393	**1.000**	0.000
	Transformational	0.930	0.990	0.840	0.360	0.887	0.720
	Situational	0.913	0.993	0.090	0.390	0.960	0.120

Table 18: Team-size = 5 ablation (3 seeds, $n = 300$ per condition). The transactional lock-in signature stays invariant (1.000 on every model) with zero exploration, while transformational and situational explore and lower lock-in. On `llama-4-scout` social (the unreliable-round-0 combination), the transformational/situational accuracy advantage over flat and transactional persists at $n = 5$, confirming the round-0-reliability finding is not an artifact of the 3-agent team size.

B.11 Round-Budget Ablation

We re-evaluate the AlphaNLI and social regimes with `round_budget`=6 (instead of the 3–4 used in the main runs) on all three models, to verify that the main results do not depend on the round-budget cap and that exploration-oriented controllers are not disadvantaged by it (Table 19; 3 seeds, $n = 300$ per condition).

B.12 Boundary probes: full results

These are the four regimes used in Section 6.1 to test where a controller can beat plain interaction, reported here in full. All use the same shared-round-0 protocol, 3-agent teams, five conditions, and three seeds as

Model	Condition	AlphaNLI	Social
gpt-oss-120b	Flat	0.913	0.427
	Transactional	0.923	0.423
	Transformational	0.903	0.440
	Situational	0.900	0.437
llama-4-scout	Flat	0.893	0.413
	Transactional	0.867	0.440
	Transformational	0.887	0.477
	Situational	0.880	**0.527**
gemma-4-31B-it	Flat	0.940	0.383
	Transactional	0.927	0.393
	Transformational	0.923	0.377
	Situational	0.927	0.383

Table 19: Round-budget = 6 ablation (3 seeds, $n = 300$ per condition), accuracy by regime. The cap is not binding, transactional still stops at round 0 (as in the main runs, Table 13), and the accuracy ranking is preserved. On `llama-4-scout` social, the one unreliable-round-0 combination, situational *gains* from the larger budget (0.527 vs. 0.513 at budget 3), consistent with reopening helping more when more incorrect round-0 majorities remain to repair.

the main matrix. ANLI-R3 (Nie et al., 2020) is the most adversarial round of Adversarial NLI (balanced three-way entailment/neutral/contradiction, $n = 100$; Table 20); Winogrande (Sakaguchi et al., 2021) is the twin-sentence coreference benchmark ($n = 100$, binary; Table 21). Both have objective gold, and the 3-agent round-0 majority stays reliable (round-0 vote 0.70–0.90) even though individual agents are often fooled, so there is little incorrect consensus to repair and no controller separates from flat. The Scruples regime (Lourie et al., 2021) is built from `r/AmITheAsshole` anecdotes: we keep only *contested* posts (plurality of community votes below 0.7 over at least five votes), balanced across a three-way verdict (author at fault / not at fault / mixed-or-unclear), $n = 99$, median community agreement 0.56 (Table 22). Here the round-0 majority is genuinely unreliable (vote 0.39–0.49) but unrecoverable: predicting the crowd's plurality on a contested moral case is not derivable from the text by reasoning, so recovery is near zero for every controller (including flat), and again no controller beats flat. Only the social regime of the main matrix combines an unreliable round-0 with a recoverable signal that flat does not already exploit (Table 7).

Condition	gpt-oss-120b	llama-4-scout	gemma-4-31B-it
Round-0 vote	0.697	0.730	0.717
Flat	0.660	0.767	0.723
Theory-free	0.683	0.707	0.737
Transactional	0.700	0.733	0.717
Transformational	0.673	0.693	0.740
Situational	0.673	0.703	0.727

Table 20: ANLI-R3 (adversarial NLI), accuracy, $n = 100$, 3 seeds. Objective gold; the round-0 majority is reliable and no controller separates from flat.

B.13 Situational trigger sensitivity

The situational controller's reopening rule uses fixed thresholds and a hand-built conflict-marker lexicon (Section 3). To test whether the social-regime gain depends on those choices, we re-run the social regime with three perturbed variants alongside the main controller, all sharing round-0 (an independent run with its own generated-once round 0, so accuracies differ from the main matrix by at most ∼2pp): *split-3-only* drops the 2–1-split trigger and reopens only on a full three-way disagreement; *lenient* relaxes the unanimous-

Condition	`gpt-oss-120b`	`llama-4-scout`	`gemma-4-31B-it`
Round-0 vote	0.843	0.783	0.900
Flat	0.860	0.787	0.913
Theory-free	0.853	0.810	0.910
Transactional	0.847	0.783	0.900
Transformational	0.857	0.780	0.897
Situational	0.843	0.780	0.893

Table 21: Winogrande (twin-sentence coreference), accuracy, $n = 100$, 3 seeds. Objective gold; the round-0 majority is reliable and no controller separates from flat.

Condition	`gpt-oss-120b`	`llama-4-scout`	`gemma-4-31B-it`
Round-0 vote	0.451	0.391	0.485
Flat	0.461	0.394	0.492
Theory-free	0.451	0.394	0.488
Transactional	0.451	0.391	0.481
Transformational	0.448	0.394	0.495
Situational	0.461	0.384	0.481

Table 22: Scruples contested moral judgments, accuracy, $n = 99$, 3 seeds. Subjective/contested gold; the round-0 majority is unreliable (vote 0.39–0.49) but unrecoverable, so every controller, including flat, stays at the vote.

but-divergent branch (rationale openings ≥ 2, conflict markers from ≥ 1 agent); and *alt-lexicon* swaps in a different but comparable conflict-marker word list. Table 23 reports accuracy, gain over flat, and recovery.

Model	**Situational variant**	Acc.	Δ vs flat	Recov.
`gpt-oss-120b`	(flat / round-0 vote 0.427)	0.437	n/a	0.070
	main	0.420	−0.017	0.047
	split-3-only	0.433	−0.003	0.023
	lenient	0.437	+0.000	0.064
	alt-lexicon	0.440	+0.003	0.064
`llama-4-scout`	(flat / round-0 vote 0.433)	0.450	n/a	0.065
	main	0.527	+0.077*	0.282
	split-3-only	0.457	+0.007	0.071
	lenient	0.540	+0.090*	0.329
	alt-lexicon	0.497	+0.047	0.218
`gemma-4-31B-it`	(flat / round-0 vote 0.380)	0.383	n/a	0.027
	main	0.400	+0.017	0.059
	split-3-only	0.390	+0.007	0.016
	lenient	0.400	+0.017	0.065
	alt-lexicon	0.410	+0.027	0.075

Table 23: Situational sensitivity on the social regime (3 seeds; independent shared-round-0 run, so both accuracy and recovery drift modestly from the main matrix, e.g. flat recovery 0.065 here vs. 0.04 there, and within-run contrasts are the comparison of interest). Δ vs flat is paired-permutation; * marks a significant gain $\geq$ 5pp. On `llama-4-scout` the gain survives the lexicon swap (+4.7pp, recovery 0.22) and *grows* under threshold relaxation (+9.0pp), so it is not an artifact of the markers or cutoffs; it collapses only under *split-3-only*, which removes the 2–1-split trigger and with it the recovery (0.28 → 0.07). `gpt-oss-120b` and `gemma-4-31B-it` show no gain under any variant.

First, the gain is *not* an artifact of the hand-built lexicon or the precise thresholds: it survives swapping the conflict-marker list and grows when the thresholds are relaxed. Second, it is mechanistically localized to

one trigger: reopening 2–1 majority splits. Removing that trigger (*split-3-only*) drops recovery from 0.28 to 0.07 and the gain to near zero, identifying reopening of thin, contested majorities (not the lexicon or the fine thresholds) as the active ingredient, which is precisely the readiness-gap intervention of Section 6.3.

C Policy Pseudocode

This appendix makes the controller logic more explicit. The goal is not to claim that the current implementations are the only valid operationalizations of these leadership theories, but to show the concrete decision flow used in the experiments. The first three subsections cover the three main controllers; the fourth, inclusive leadership, is the dissent-preservation probe of Appendix B.4, not an additional main controller.

C.1 Transactional Leadership

```
Input:  task x, round-t answers Y_t
m ← majority answer among non-empty Y_t
for each agent response y_{i,t}:
  if y_{i,t} empty or y_{i,t} ≠ m:  issue REVISE
  else (y_{i,t} = m):  issue ACCEPT
if majority count ≥ 2 (2-of-3):
  SYNTHESIZE final answer from accepted pool
else:
  continue to next round
```

The acceptance criterion is concrete: an agent's answer is accepted if it matches the current-round majority answer and revised otherwise; the team closes the round on a 2-of-3 majority. The defining property is strong convergence pressure: disagreement is treated primarily as a signal for correction rather than for further exploration.

C.2 Transformational Leadership

```
Input:  task x, round-t answers Y_t, team goal g_t
broadcast shared objective g_t to the team
for each agent:
  assign differentiated exploration directive d_{i,t}
  request alternative framing, evidence path, or hypothesis
if distinct plausible lines remain active:
  delay synthesis and preserve diversity one more round
else:
  SYNTHESIZE final answer
```

The defining property is structured diversity around a common team objective rather than local correction toward immediate agreement.

C.3 Situational Leadership

```
Input:  task x, round-t answers Y_t (3 agents)
u ← # distinct non-empty answers
d ← # distinct rationale openings (first-sentence prefixes)
c ← # agents whose text contains a conflict marker
if u ≥ 3 (three-way split):  open EXPLORE
elif u = 2 (2-1 split):  open EXPLORE
elif u = 1 and d ≥ 3 and c ≥ 2:  open EXPLORE
else:  fall back to transactional-style convergence
  (REVISE / ACCEPT, then SYNTHESIZE on a 2-of-3 majority)
```

The conflict markers are a fixed lexicon ("however," "but," "although," "contradict," "unlikely," "fails," and similar). The defining property is that switching depends on interaction state rather than benchmark identity. The controller does not need to know whether a task came from AlphaNLI, Social Chemistry, or closed-ended QA; it reacts to whether the team currently looks prematurely aligned or genuinely resolved.

C.4 Inclusive Leadership

```
Input:  task x, round-t answers Y_t
identify current majority and minority positions
if a minority objection exists:
  issue OBJECT to the minority side
  issue DEFEND to the majority side
  preserve the objection in the interaction record
  RECONSIDER before final synthesis
else:
  proceed directly to synthesis
```

The defining property is not generic delay, but explicit preservation of minority dissent before convergence.