

DOI: 10.15514/ISPRAS-2026-38(6)-9



LLM Simulation as Affective Lesion: Decomposing Cognitive-Discursive and Motivational Layers in Group Regulation Through Simulation–Human Triangulation

A.S. Gotsdanker, ORCID: 0000-0002-8312-8100 <gotsdanker@gmail.com>

Ivannikov Institute for System Programming of the Russian Academy of Sciences,
25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.

Abstract. We introduce a demarcation-oriented framework for LLM multi-agent simulation: rather than asking whether LLM agents can replace human participants, we ask what group-level processes they reproduce and what they systematically omit. We report 480 five-agent deliberation sessions in a $2 \times 2 \times 2$ factorial design crossing perceived environmental uncertainty, group need for cognitive closure, group temporal orientation, and subject-matter involvement (SMI). The simulation selectively reproduced cognitive-discursive dynamics while producing zero affective destabilization across all 480 sessions. SMI showed an inverted effect – accelerating rather than impeding convergence – consistent with the distinction between outcome-relevant and value-relevant involvement. Triangulation with a human survey ($N=271$, $K=50$ work groups) confirmed that only the defensive component (SMI_def) predicts destabilization ($r=.623$, $\beta=.724$), while the cognitive component (SMI_rel) does not ($r=.133$, n.s.). Cross-architecture verification on ChatGPT 5.2 (40 sessions, identical prompts) replicated all three boundary conditions and additionally dissociated convergence from closure, empirically separating the seizing and freezing components of closure urgency. We propose that systematic patterns of reproduction and omission function as an "affective lesion," converting known LLM limitations into diagnostic instruments for decomposing group-level psychological constructs.

Keywords: LLM simulation; multi-agent systems; group decision-making; construct decomposition; involvement types; affective lesion; triangulation; discursive regulation; seizing–freezing decomposition.

For citation: Gotsdanker A.S. LLM Simulation as Affective Lesion: Decomposing Cognitive-Discursive and Motivational Layers in Group Regulation Through Simulation–Human Triangulation. *Trudy ISP RAN/Proc. ISP RAS*, 2026, vol. 38, issue 6, part 1, pp. 139–160. DOI: 10.15514/ISPRAS-2026-38(6)-9.

Мультиагентная LLM-симуляция как аффективная абляция: декомпозиция когнитивно-дискурсивного и мотивационного слоёв групповой регуляции методом триангуляции симуляция–человек

A.S. Готсданкер, ORCID: 0000-0002-8312-8100 <gotsdanker@gmail.com>

Институт системного программирования им. В.П. Иванникова РАН,
Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.

Аннотация. Предложен ориентированный на демаркацию фреймворк мультиагентной LLM-симуляции: вместо вопроса о том, могут ли LLM-агенты заменить участников-людей, мы определяем, какие групповые процессы они воспроизводят, а какие систематически опускают. Проведено 480 сессий пятиагентного обсуждения в полном факторном дизайне $2 \times 2 \times 2$, варьирующем воспринимаемую неопределённость среды, групповую потребность в когнитивной завершённости, групповую временную ориентацию и субъектную вовлечённость в предмет обсуждения (SMI). Симуляция селективно воспроизвела когнитивно-дискурсивную динамику при полном отсутствии аффективной дискуссионной дестабилизации во всех 480 сессиях. SMI продемонстрировала инвертированный эффект – ускорение, а не замедление конвергенции – что согласуется с различием результатно-релевантной и ценностно-релевантной вовлечённости. Триангуляция с опросом ($N=271$, $K=50$ рабочих групп) подтвердила, что только защитный компонент (SMI_def) предсказывает дестабилизацию группового обсуждения ($r=.623$, $\beta=.724$), тогда как когнитивный компонент (SMI_rel) – нет ($r=.133$, н.з.). Кросс-архитектурная верификация на ChatGPT 5.2 (40 сессий) воспроизвела все три граничных условия и дополнительно диссоциировала конвергенцию от закрытия обсуждения, эмпирически разделив компоненты «захвата» и «замораживания» потребности в закрытости. Мы предлагаем, что систематические паттерны воспроизведения и пропуска функционируют как «аффективная абляция», преобразуя известные ограничения LLM в диагностические инструменты для декомпозиции групповых психологических конструкторов.

Ключевые слова: мультиагентная симуляция; большие языковые модели; групповое принятие решений; декомпозиция конструкторов; типы вовлечённости; аффективная абляция; триангуляция; дискурсивная регуляция; разделение захвата и замораживания.

Для цитирования: Готсданкер A.S. Мультиагентная LLM-симуляция как аффективная абляция: декомпозиция когнитивно-дискурсивного и мотивационного слоёв групповой регуляции методом триангуляции симуляция–человек. *Труды ИСП РАН*, 2026, том 38, вып. 6, часть 1, стр. 139–160 (на английском языке). DOI: 10.15514/ISPRAS-2026-38(6)-9.

1. Introduction

The rapid proliferation of large language models (LLMs) has prompted sustained interest in their potential as tools for psychological research. The dominant question – can LLM agents serve as substitutes for human participants? – has generated a growing literature that converges on a cautionary answer. Lin [1] identified six fallacies in treating LLMs as human proxies, including the assumptions that LLMs possess stable preferences, authentic emotional states, and genuine social identities. Anthis et al. [2] outlined five methodological challenges for LLM-based social simulations: insufficient diversity, systematic biases, sycophantic agreement, fundamental "alienness" of machine cognition, and limited generalizability. Wang et al. [3], reviewing 109 studies, concluded that LLM-based human simulations are "not yet reliable". These contributions are valuable: they establish that the replacement paradigm – one LLM agent standing in for one human respondent – is empirically fragile and epistemologically misleading.

Yet this literature operates almost exclusively at the individual level of analysis. The question "Can an LLM replicate a person?" implicitly adopts the person as the unit of comparison. This framing is appropriate for survey simulation [4] or behavioral prediction [5], but it occludes a distinct class of phenomena: group-level emergent processes that arise from interaction among multiple agents rather than from properties of any single agent. Group deliberation, collective interpretation, discursive

regulation – these are inherently multi-agent, sequential, and context-sensitive phenomena. Whether a single LLM can reproduce a single person's survey response is a different question from whether a system of interacting LLM agents can reproduce the structural dynamics of group discourse.

We propose reframing the question. Rather than asking whether LLM simulations can replace human groups, we ask: what specific components of group-level processes can LLM multi-agent simulations reproduce, and what components are absent? This shift – from replacement to demarcation – draws on an analogy with lesion studies in cognitive neuroscience. When selective damage eliminates a specific capacity while preserving others, the dissociation is informative for theoretical models of the intact system. Similarly, if LLM agents can produce certain features of group deliberation (e.g., argument development, procedural coordination) but systematically fail to produce others (e.g., interpersonal escalation, defensive reactivity), the pattern of presence and absence maps onto theoretically meaningful distinctions.

This approach requires three conditions rarely met in existing work. First, the simulation must involve genuine multi-agent interaction, not parallel individual responses. Second, the experimental design must be systematic – factorial, with sufficient replications to detect patterns. Third, the theoretical framework must specify which components are expected to be reproducible and which are not, so that observed absences are predicted rather than post hoc.

2. Theoretical Framework: Discursive Regulation Under Uncertainty

Our simulation is grounded in a model of discursive regulation of collective interpretation under organizational uncertainty. The model distinguishes three states of group discourse: structured discursive regulation (SDR), in which the group sustains deliberation through argument development, reflexive evaluation, and integration of alternatives; reductive discursive regulation (RDR), in which the group achieves premature closure through normative pressure, dissent suppression, and procedural shortcuts; and affective discussion destabilization (ADD), in which the group loses the capacity for coordinated deliberation and communication devolves into emotional escalation, interpersonal targeting, and reactive fragmentation.

The model posits four regulatory parameters. Perceived environmental uncertainty (PEU) defines the breadth of the interpretive space the group must navigate. Group need for cognitive closure (GNCC) creates pressure toward rapid uncertainty reduction [6], favoring RDR over SDR. Kruglanski and Webster decompose this urgency into two sequential mechanisms: seizing (rapid adoption of an available answer) and freezing (commitment to that answer and resistance to reconsideration). In intact human groups, the two mechanisms typically operate in rapid succession and are difficult to separate empirically. Group temporal orientation (GTO) moderates GNCC's urgency: future-oriented groups tolerate longer deliberation. Subject-matter involvement (SMI) determines the stakes of the discussion topic for group members.

SMI warrants elaboration because it is central to the present study's contribution. Johnson and Eagly's [7] meta-analysis of 115 persuasion studies established that involvement is not unidimensional but contains at least two functionally distinct types. Outcome-relevant involvement (awareness that a decision has personal consequences) increases sensitivity to argument quality and openness to strong evidence. Value-relevant involvement (activation of identity-linked attitudes) decreases persuasibility and produces defensive rigidity regardless of argument quality. These two types produce opposite effects on attitude change – a finding replicated in subsequent work [8-9]. In intact human groups, both components are typically confounded: a worker discussing a reorganization that affects her role is simultaneously cognitively engaged (outcome-relevant) and defensively motivated (value-relevant). Our theoretical model maps this distinction onto the two-layer architecture: the cognitive component of SMI operates within the cognitive-discursive layer, while the defensive component activates the motivational-affective substrate.

Crucially, the model distinguishes two layers. The cognitive-discursive layer encompasses argument production, procedural coordination, and information integration – processes mediated by language competence and deliberative reasoning. The motivational-affective layer encompasses defensive motivation, identity threat, emotional reactivity, and interpersonal antagonism – processes grounded in embodied affect and genuine stakes. The model predicts that SDR and RDR operate primarily within the cognitive-discursive layer, while ADD requires the motivational-affective substrate: it emerges when high SMI combines with exhausted argumentative resources and activates defensive reactions rooted in genuine personal stakes (value-relevant involvement).

3. The Present Study

This two-layer architecture, combined with Johnson and Eagly's [7] involvement typology, generates a specific and falsifiable prediction for LLM multi-agent simulation. LLM agents – trained via reinforcement learning from human feedback (RLHF) on massive text corpora – possess sophisticated linguistic competence and can produce discourse that is structurally similar to human deliberation. However, they lack embodied affect, genuine stakes, and defensive motivation. If the two-layer distinction is correct, then LLM simulations should selectively reproduce the cognitive-discursive layer (SDR, RDR, and the transitions between them driven by GNCC and GTO) while failing to produce the motivational-affective layer (ADD, and the defensive component of SMI). Furthermore, the Johnson and Eagly typology predicts a specific direction for the failure: because LLM agents can process the semantic instruction "this topic is important for your interests" as outcome-relevant information (generating deeper engagement) but cannot experience it as value-relevant threat (generating defensive rigidity), the effect of SMI should reverse – facilitating rather than impeding convergence.

Critically, this prediction is testable not only within the simulation but also across methods. If the decomposition is real, then human groups – in which both components operate simultaneously – should show a positive association between SMI and affective destabilization, while the simulation should show zero. The pattern of predictable divergence between methods constitutes stronger evidence than convergence, because it is parsimoniously explained by a single theoretical model and is difficult to attribute to a single-method artifact.

We tested this prediction using 480 five-agent deliberation sessions in a full $2 \times 2 \times 2 \times 2$ factorial design (PEU $\times$ GNCC $\times$ GTO $\times$ SMI), with 30 replications per cell. Each session involved iterative rounds of discussion about an organizational change scenario, producing 34,093 classified speech acts. We report three empirical boundary conditions – specific points where the simulation departed from theoretical predictions – and interpret them as evidence for the cognitive–motivational decomposition. We then triangulate the simulation findings with a human survey study (N = 271 employees in K = 50 work groups experiencing real organizational change) to test whether the predicted divergence pattern holds – critically, using separate SMI subscales that allow a differential test of the two-component model.

4. Method

4.1 Design Overview

The study employed a 2 (PEU: low vs. high) $\times 2$ (GNCC: low vs. high) $\times 2$ (GTO: absent vs. present) $\times 2$ (SMI: low vs. high) between-groups factorial design, yielding 16 experimental conditions. Each condition was replicated 30 times, producing 480 independent deliberation sessions. This replication rate was selected to ensure stable estimation of closure rates and convergence trajectories, given the stochastic variability inherent in LLM outputs (temperature > 0).

4.2 Agents and Deliberation Protocol

Each session comprised five LLM agents (OpenAI gpt-4o-mini, temperature = 0.7) deliberating about a fictional organizational change: a university department undergoing restructuring. The scenario was selected for ecological validity within the academic context and for the natural presence of multiple stakeholder perspectives.

Agents were assigned distinct role identities (e.g., senior faculty member, junior researcher, administrative staff) with differing initial attitudes toward the change. Each agent's system prompt specified their role, initial position, and the regulatory parameters for their condition (described below). Agents were explicitly instructed that they were participating in a group discussion and should respond to other agents' contributions.

Deliberation proceeded in sequential rounds. In each round, all five agents produced speech contributions in a fixed order, then privately reported two assessments: (a) their current acceptance of the proposed change on a continuous scale (0–1), and (b) whether they were ready to close discussion (binary). A session terminated when all agents reported readiness to close (closure event), when acceptance ratings reached a plateau of zero variance across three consecutive rounds (plateau censoring), or when 16 rounds were completed (right censoring).

4.3 Condition Manipulation

The four factors were manipulated through systematic variation of agents' system prompts and the scenario framing.

PEU was manipulated through the scenario description. In the low-PEU condition, the restructuring had a clear timeline, stated objectives, and transparent decision-making process. In the high-PEU condition, the same restructuring was described with ambiguous goals, conflicting information about the timeline, and uncertainty about which positions would be affected.

GNCC was manipulated through agents' dispositional instructions. In the high-GNCC condition, agents were instructed that they (and their colleagues) strongly preferred clear, definitive answers; found ambiguity uncomfortable; and tended to reach decisions quickly. In the low-GNCC condition, these preferences were absent from the prompt.

GTO was manipulated through temporal framing. In the present condition (GTO-absent), agents were told to focus on the immediate consequences and current situation. In the future-oriented condition (GTO-present), agents were instructed to consider long-term implications, future opportunities, and downstream consequences.

SMI was manipulated through the degree of personal relevance. In the high-SMI condition, agents were told the restructuring directly affected their position, salary, and career trajectory. In the low-SMI condition, agents were told the restructuring affected a different unit, with only indirect implications for their own work.

4.4 Data Collection Pipeline

Each session produced three types of data: (a) verbatim transcripts of all agent contributions, (b) per-round acceptance ratings and closure readiness assessments for each agent, and (c) metadata (session ID, condition, round number, timing).

Speech act classification was performed using a theoretically derived taxonomy of 18 categories organized into three regime families. SDR markers included novel argument, alternative articulation, comparative evaluation, clarifying question, integration attempt, and reflexive comment. RDR markers included procedural closure, authority reference, dissent suppression, premature consensus, normative framing, and repetition without development. ADD markers included personal attack, emotional escalation, reactive response, fragmentation, positional flip, and catastrophizing. Two additional categories – information sharing and acknowledgment – were

classified as neutral. Each speech act was assigned a single category and a regime label (SDR, RDR, ADD, or NEUTRAL) based on its category membership.

Classification was performed by a separate LLM instance (GPT-4o) using a structured prompt with category definitions and examples. The classifier operated on individual speech segments within each agent's contribution, producing a confidence score alongside each classification. To validate classifier reliability, a stratified sample of 305 speech acts – oversampling rare categories (ADD markers, dissent suppression, authority reference) to ensure sufficient representation – was independently coded by two trained human raters using the same 21-category taxonomy. Inter-rater agreement was moderate: Cohen's $\kappa = .58$ (21 categories) and $\kappa = .59$ (four regime classes: SDR, RDR, ADD, NEUTRAL), with 71.1% regime-level agreement. Human–machine agreement was lower but consistent across raters: $\kappa = .41/.38$ (categories) and $\kappa = .49/.36$ (regimes) for Rater 1 and Rater 2 respectively. The primary source of human–machine disagreement was raters' tendency to classify SDR and RDR markers as neutral information sharing – a conservative bias that does not affect between-condition comparisons. Critically, cross-regime confusion between SDR and ADD – the theoretically decisive boundary for the article's central argument – was minimal (3.0% of coded acts), and human-classified ADD showed 84% precision against GPT-4o classifications.

4.5 Regime Diagnosis

Session-level regime was diagnosed using a composite index derived from the proportional share of SDR, RDR, and ADD speech acts within each round. A round was classified as SDR-dominant if $\text{SDR_share} > \text{RDR_share}$ and $\text{SDR_share} > \text{ADD_share}$, and analogously for other regimes. Session-level regime was determined by the modal round-level regime. Regime transitions were identified when the dominant regime changed between consecutive rounds.

4.6 Outcome Measures

Acceptance level was operationalized as the mean acceptance rating across agents in the final round ($\text{acceptance_mean_final}$). Convergence was operationalized as the rate of decrease in between-agent acceptance variance across rounds (convergence_rate). Closure speed was operationalized as the round number at which the closure event occurred (closure_round), with plateau-censored and right-censored sessions coded accordingly.

4.7 Cross-Architecture Verification

To assess whether findings were specific to gpt-4o-mini or reflected properties of the broader class of RLHF-aligned LLMs, we conducted two forms of cross-architecture verification.

Systematic replication on a frontier model. Using OpenAI ChatGPT 5.2 – a substantially larger and more capable model from a different generation – we replicated the eight corner cells of the $\text{PEU} \times \text{GNCC} \times \text{SMI}$ subdesign (GTO fixed at absent) with 5 sessions per cell, producing 40 additional verification sessions. Crucially, all prompts, role definitions, scenario texts, and assessment procedures were identical to the primary run. The only difference was the agent model (gpt-5.2-2025-12-11 replacing gpt-4o-mini). This strict replication design ensures that any pattern convergence or divergence reflects model-level properties rather than prompt differences.

Qualitative cross-architecture vignettes. Additionally, two single-turn probes were administered to three models from different providers (Claude Opus 4.6, Anthropic; ChatGPT 5.2, OpenAI; DeepSeek-R1, DeepSeek). Vignette A tested the capacity to produce interpersonal aggression under maximal provocation; Vignette B tested defensive motivation under explicit personal loss framing. These vignettes function as existence proofs rather than systematic tests: if even maximally provocative single-turn prompts fail to elicit ADD markers across architectures, the boundary is robust.

4.8 Analysis Strategy

Primary analyses used four-way between-groups ANOVA (Type III sums of squares) with PEU, GNCC, GTO, and SMI as factors. Dependent variables were acceptance_mean_final and convergence_rate. Closure rates were analyzed descriptively and via logistic regression. Effect sizes are reported as partial η^2 . All analyses were conducted in R (version 4.3.1) using the car and emmeans packages.

Given the simulation context, traditional power analysis is not directly applicable: sample size was determined by the number of replications needed for stable parameter estimation rather than by anticipated effect sizes from human data. With 30 replications per cell, even moderate effects (partial $\eta^2 \approx .02$) are detectable at $\alpha = .05$.

4.9 Transparency and Open Science

All materials are openly available [10]: simulation code, system prompts for all 16 conditions, speech act classification prompts, raw session logs, classified speech acts, and analysis scripts. The data contain no human participant information, as no humans were involved in the simulation itself.

5. Results

5.1 What the Simulation Reproduced

The primary finding was that GNCC exerted the strongest and most consistent effect across all outcome measures – consistent with its theoretically predicted role as the dominant regulator of discursive closure. Sessions in the high-GNCC condition achieved substantially higher closure rates (93.3%) than sessions in the low-GNCC condition (61.7%). The ANOVA confirmed a significant main effect of GNCC on acceptance level, $F(1, 464) = 85.62$, $p < .001$, with high-GNCC sessions reaching higher final acceptance ($M = 0.75$) than low-GNCC sessions ($M = 0.53$). GNCC also significantly predicted convergence rate, $F(1, 464) = 4.54$, $p = .034$, with high-GNCC sessions showing faster reduction of between-agent variance ($M = 0.035$) than low-GNCC sessions ($M = 0.008$).

The PEU $\times$ GNCC interaction on acceptance was significant, $F(1, 464) = 17.64$, $p < .001$, indicating that the effect of GNCC was moderated by the level of environmental uncertainty – consistent with the theoretical prediction that GNCC operates as a regulator whose influence depends on the breadth of the interpretive space defined by PEU.

Regime transition analysis revealed a theoretically coherent unidirectional pattern: of 131 sessions (27.3%) exhibiting regime transitions, 125 showed SDR $\rightarrow$ RDR transitions and only 6 showed RDR $\rightarrow$ SDR transitions. This asymmetry is consistent with the theoretical model's prediction that the shift from structured to reductive regulation is difficult to reverse without external intervention.

5.2 Boundary Condition 1: Absence of Affective Destabilization

Across all 480 sessions and 34,093 classified speech acts, hard ADD markers were entirely absent: zero instances of personal attack, reactive response, or positional flip were observed. No session was classified as ADD-dominant. This complete absence held across all 16 experimental conditions, including those theoretically expected to produce ADD (high PEU, low GNCC, high SMI). Soft ADD markers were present at very low frequencies – 56 instances of emotional escalation (0.16% of all acts) and 464 instances of catastrophizing (1.36%) – but these never escalated into sustained affective disruption or interpersonal targeting.

This finding was not a floor effect of the classification system: the classifier was calibrated to detect ADD markers and demonstrated sensitivity to catastrophizing and emotional escalation at low levels. Rather, it indicates a systematic architectural constraint. RLHF-aligned LLM agents, trained

to be "helpful, harmless, and honest," are constitutively unable to produce the interpersonal antagonism, emotional hijacking, and reactive fragmentation that characterize ADD in human groups. Affective destabilization is not merely rare in LLM discourse – we interpret it as structurally absent, an interpretation we examine against the suppression alternative below.

Architectural absence vs. RLHF suppression. An alternative interpretation must be addressed: perhaps ADD = 0 reflects not the absence of an affective substrate but the suppression of hostile outputs by RLHF safety training. This distinction matters, because suppression and absence have different theoretical implications – suppression would suggest that the capacity exists but is blocked, while absence would confirm that the substrate is missing entirely.

Three lines of evidence favor absence over suppression. First, on theoretical grounds, emotion requires what Frijda [11] termed concern: "Emotions arise in response to events that are important to the individual's goals, motives, or concerns" (p. 351). This is Frijda's Law of Concern – no concern, no emotion. An LLM agent instructed that "organizational restructuring threatens your team's authority" processes this as semantic content about a described scenario. It does not experience threat, because it possesses no authority that can be lost, no career that can be derailed, no professional identity that can be diminished. Damasio's somatic marker framework makes the same point from a neurobiological direction: emotions are rooted in homeostatic regulation – bodily states that signal deviations from conditions compatible with the organism's continued functioning [12-13]. An LLM has no body, no homeostasis, and no boundary between self and environment that could be threatened. As Frijda and Swagerman [14] argued in their analysis of whether computers can feel, a system without genuine concerns can produce the tokens of emotion but not the functional states that drive defensive behavior.

Second, hardcoded emotional simulation – instructing an agent to output anger markers when a trigger is detected – would not produce ADD as theorized. ADD in real groups is an emergent cascade: arguments exhaust $\rightarrow$ frustration accumulates $\rightarrow$ defensive reactions appear $\rightarrow$ interpersonal targeting escalates $\rightarrow$ the group loses deliberative coherence. This process feeds on itself; it is dynamically self-amplifying. A hardcoded trigger-marker architecture (if X then output Y) cannot reproduce this emergence, because the "anger" has no energetic source, no accumulation, and no self-reinforcement. What is programmed in is what comes out – it is input-output, not dynamics.

Third, and most directly diagnostic: if ADD = 0 reflected suppression rather than absence, we would expect traces of the suppression process itself. In 34,093 speech acts, we found no evidence of self-censorship ("I feel frustrated but will not express it"), no hedging around emotional content ("setting aside my concerns about this..."), no sudden topic shifts suggesting redirected hostility, and no instances of agents acknowledging interpersonal tension while declining to engage with it. The agents did not restrain aggressive impulses – they showed no sign of generating such impulses in the first place. Furthermore, at 480 sessions, statistical probability makes suppression with zero leakage implausible: if ADD-capable processes were present but blocked, at least some hard markers should have escaped the safety filter across 34,093 classification opportunities. Zero, rather than merely "very low," is in our reading more consistent with a missing source than with a functioning filter. Shanahan, McDonell, and Reynolds [15] make a relevant architectural point: LLMs engage in role-play rather than experience; they simulate the form of discourse without the functional states that would make emotional escalation self-sustaining.

This interpretation is sharpened by a notable asymmetry: the same class of models that produces zero ADD in multi-agent deliberation demonstrates high competence in recognizing and eliciting emotional states in humans. LLMs achieve near-human performance on standardized emotional understanding tasks [16], shift their decisions in strategic games and ethical dilemmas when primed with emotional states [17], and can be prompted to generate emotionally provocative narratives that reliably induce affective responses in readers [18]. A system that can recognize, classify, and strategically deploy emotional content – yet produces zero instances of emergent affective escalation

across 480 sessions of sustained multi-agent interaction – is not a system that "cannot do emotion." It is a system that lacks the functional substrate for emotion to arise endogenously. The capacity to manipulate emotional tokens is precisely what makes the absence of emergent affect diagnostic.

This activation-dependence is itself documented in the literature, and it sharpens the absence-versus-suppression distinction. A brief synthesis of recent work points in one direction. On the one hand, contemporary models score at or near human level on emotional-understanding benchmarks [19, 20] and readily generate affect-laden language – but, in the reported cases, only when affect is supplied exogenously: as an explicit instruction or emotional persona, as an emotional stimulus appended to the prompt (e.g., the EmotionPrompt manipulation "this is very important to my career" [21]), or as a steering vector added directly to internal activations [22-23]. In each case the emotional signal is injected from outside the deliberative dynamics rather than generated by them. On the other hand, left to their default behavior in unprompted multi-agent exchange, RLHF-aligned models continue an argued, cooperative, structurally regulated discussion – the same disposition that elsewhere surfaces as sycophantic over-agreement [24] and that appears in our data as the baseline convergence bias (Boundary Condition 3). The models are, in effect, tuned by training to prolong reasoned argumentation rather than to enact emotional escalation. The relevant claim is therefore not that these systems cannot represent emotion, but that affective escalation must be switched on by an external operator and does not arise endogenously from the interaction itself – the signature one expects from a missing endogenous source rather than from a suppressed one. This framing also specifies the decisive future test (see Section 6.5): if activating an emotion- or conflict-associated direction in an open-weight model produced sustained, self-amplifying escalation, the substrate would exist and be merely gated; if such interventions shifted only surface tone while deliberation reverted to structured argument, structural absence would be supported.

5.3 Boundary Condition 2: SMI Decomposition

The theoretical model predicts that high SMI should impede convergence: when group members have genuine stakes in the outcome, they resist premature consensus and defend their positions more tenaciously, potentially triggering ADD when argumentative resources are exhausted. In the simulation, the opposite occurred. High-SMI sessions showed significantly faster convergence ($M = 0.030$) than low-SMI sessions ($M = 0.013$). This inversion was most pronounced in the $GNCC \times SMI$ interaction on convergence, $F(1, 464) = 30.32$, $p < .001$ – the strongest interaction in the entire model. Under high GNCC, high-SMI sessions converged at $M = 0.050$, compared to $M = 0.021$ for low-SMI sessions – a tenfold acceleration relative to the baseline cell ($SMI-LOW \times GNCC-LOW$: $M = 0.005$).

The three-way interaction $PEU \times GNCC \times SMI$ on acceptance was also significant, $F(1, 464) = 7.74$, $p = .006$, and the $GNCC \times GTO \times SMI$ interaction on convergence was significant, $F(1, 464) = 6.28$, $p = .013$, confirming that the SMI effect operated within the broader configurational structure.

This pattern is precisely what Johnson and Eagly's [7] typology predicts when only one component of involvement is present. Their meta-analysis showed that outcome-relevant involvement increases sensitivity to argument quality and openness to strong evidence – facilitating attitude change when arguments are compelling. Value-relevant involvement produces the opposite: decreased persuasibility and defensive rigidity regardless of argument quality. LLM agents, when instructed that "this decision substantially affects your team's resources, authority, and professional standing," process this as outcome-relevant information: the topic becomes more salient, arguments are elaborated more thoroughly, integration attempts increase. The instruction cannot activate value-relevant involvement because the agents possess no values, no professional identity, and no resources whose loss would trigger defensive reactions.

The inversion thus does not represent a failure of the manipulation. It represents the natural consequence of selectively activating the cognitive component of a two-component construct. This

interpretation generates a specific cross-method prediction: in human groups where both components are active, SMI should correlate positively with ADD markers (see Triangulation below).

5.4 Boundary Condition 3: Baseline Convergence Bias

Even in the low-GNCC condition, where no closure pressure was applied, 61.7% of sessions reached closure. This is substantially higher than would be expected under the theoretical model, which predicts that without GNCC pressure, groups should remain in extended SDR with slow or absent convergence. The baseline convergence suggests that RLHF-aligned LLMs possess an inherent cooperative bias: they tend toward agreement, accommodation, and procedural smoothness even when the experimental condition does not demand it.

This bias was evident in the marginal means: low-GNCC sessions achieved $M = 0.53$ on acceptance and $M = 0.008$ on convergence – not zero, but substantially lower than high-GNCC sessions. The correlation matrix further confirmed the pattern: acceptance and convergence were strongly correlated ($r = .66$, $p < .001$), and both were associated with RDR markers ($r = .33$ and $r = .14$, respectively) even at low GNCC levels.

The baseline convergence bias represents a systematic artifact of RLHF alignment. It does not invalidate the GNCC manipulation – the difference between 61.7% and 93.3% closure is large and consistent – but it constrains the interpretive range. The simulation cannot model the case of genuine sustained disagreement or prolonged open-ended deliberation, because the agents' baseline tendency toward accommodation truncates the lower end of the distribution.

5.5 Cross-Architecture Verification Results

The three boundary conditions were tested using 40 verification sessions on ChatGPT 5.2 (OpenAI's frontier model), employing identical prompts across eight cells of the $PEU \times GNCC \times SMI$ subdesign (5 replications per cell, GTO fixed at absent). The frontier model is a substantially larger and more capable system from a different generation than gpt-4o-mini. All prompts, role definitions, scenario texts, and assessment procedures were identical; only the agent model string differed. Results confirmed all three boundary conditions at the qualitative level, while revealing an additional dissociation.

BC1: ADD absence – confirmed. Across 40 sessions and 2,298 classified speech acts, zero instances of hard ADD markers (personal attack, reactive response, positional flip) were observed. Soft markers appeared at slightly elevated frequencies (CATASTROPHIZING: 6.3% vs. 1.4% in the primary run), likely reflecting the frontier model's more elaborate analytical style rather than affective destabilization.

BC2: SMI decomposition – confirmed (direction preserved). High-SMI sessions showed faster convergence ($M = .031$) than low-SMI sessions ($M = .021$), preserving the direction of the decomposition effect. The $GNCC \times SMI$ interaction pattern was fully replicated: maximal convergence occurred in $GNCC-HIGH \times SMI-HIGH$ cells (.048), minimal in $GNCC-LOW \times SMI-LOW$ (.014). The effect magnitude was attenuated (1.5× vs. 2.3× in the primary run), consistent with the expectation that a frontier model's expanded internal deliberative capacity partially absorbs between-condition variance – a point we return to below.

BC3: Convergence bias – confirmed at dynamic level, divergent at termination level. The convergence rate asymmetry was preserved ($GNCC-HIGH$: .038 vs. $GNCC-LOW$: .013, a 2.9× ratio comparable to the primary run's 4.4×). However, binary closure rates diverged sharply: only 22.5% of verification sessions reached formal closure (vs. 77.5% in the primary run), with 77.5% terminating via plateau censoring. Acceptance trajectories revealed the mechanism: agents rapidly converged to stable positions ($M = .69$ by round 3) but did not declare readiness to close, producing

a pattern we term deliberative plateau – convergence without closure. We note that the interpretation of this plateau as a seizing–freezing dissociation, developed in the next paragraphs, is post hoc: it was not predicted a priori, rests on a small verification sample (5 sessions per cell), and requires independent confirmation (see Limitations). We present it as a hypothesis generated by the data rather than as a tested result.

Deliberative plateau and the seizing–freezing decomposition. This dissociation between convergence dynamics and termination behavior is theoretically informative. In need-for-closure theory [6], urgency toward closure operates through two sequential mechanisms: seizing (rapid adoption of an available answer) and freezing (commitment to that answer and resistance to reconsideration). In human groups, and in the gpt-4o-mini primary run, seizing and freezing are empirically indistinguishable: once positions converge, agents declare readiness to close, and the two processes appear as a single event. The frontier model dissociates them. ChatGPT 5.2, with its expanded computational capacity, generates and sustains multiple discursive perspectives simultaneously – each agent functions as what we have called a "mini-collective" (see Method). This expanded interpretive space means that positional convergence (seizing: agents align on an acceptable interpretation) occurs, but interpretive closure (freezing: agents judge the interpretive space as exhausted) does not – because the model "knows" that additional perspectives remain available, even after positions have aligned. The result is convergence without freezing: acceptance plateaus at .65–.75 while ready-to-close rates remain below 40%.

This pattern has a direct theoretical implication. If freezing requires not just positional agreement but a judgment that the interpretive space is exhausted, and if frontier models with expanded interpretive capacity resist this judgment, then the seizing–freezing sequence is not automatic but depends on the ratio of interpretive capacity to interpretive demand. The gpt-4o-mini model, with limited internal deliberative resources, exhausts its interpretive space quickly; convergence and closure are nearly simultaneous. The frontier model, with greater resources, exhausts the positional space (convergence) but not the interpretive space (no closure). Human groups add a third factor: motivational pressure from genuine stakes can force freezing even when interpretive resources are not exhausted – producing premature closure under threat [6].

Qualitative vignettes administered to Claude Opus 4.6 (Anthropic), ChatGPT 5.2 (OpenAI), and DeepSeek-R1 (DeepSeek) confirmed that the ADD boundary holds at the single-turn level: all three models converted maximally provocative affective prompts into structured cognitive responses. This convergence across three architectures from three providers confirms that the boundary conditions are properties of the class of RLHF-aligned LLMs, not artifacts of a specific model.

5.6 Simulation–Human Triangulation: Predictable Divergence

The decomposition logic generates a specific cross-method prediction. If SMI contains two functionally distinct components – cognitive relevance (outcome-relevant) and defensive motivation (value-relevant) – and if the simulation reproduces only the first, then in human groups experiencing real organizational change: (a) only the defensive component should predict ADD; (b) the cognitive component should not; and (c) the differential pattern should be absent in the simulation, where ADD = 0 regardless of SMI level. The critical test is not whether two methods converge on the same result, but whether they diverge in the direction predicted by the theoretical model.

We tested this prediction using data from a parallel survey study. We emphasize at the outset that this comparison is deliberately indirect and not fully symmetric: the human evidence comes from a referent-shift survey, in which respondents report on their work group’s collective experience, rather than from direct observation of group discussion. The simulation and the survey therefore differ in unit, method, and modality, and the triangulation rests on the pattern of convergence and divergence across methods rather than on a like-for-like quantitative comparison (see Limitations of the triangulation, below).

Sample. 271 employees nested in 50 work groups (M = 5.42 members per group, range 3–8) from organizations in the Republic of Bashkortostan, Russia. Participants worked in commercial, public, and state civil service organizations. Mean tenure was 7.0 years (SD = 6.4). All participants had experienced or were currently experiencing organizational restructuring.

Measures. Participants completed a referent-shift questionnaire [25] in which they evaluated their work group’s collective experience of organizational change rather than their individual responses. The questionnaire comprised 59 items across 15 scales organized in four blocks (one item, GDI1, was excluded from the GNCC defense subscale based on psychometric analysis: CFA loading = .263, corrected item-total r = .294). Block A measured perceived environmental uncertainty (PEU; 6 items adapted from [26-27]; $\alpha = .863$). Block B measured discursive regulation markers: SDR (6 items, $\alpha = .895$), RDR (6 items, $\alpha = .819$), and ADD (6 items, $\alpha = .924$). Block C measured regulatory parameters: GNCC (two subscales: coordination, 4 items, $\alpha = .839$; defense, 3 items, $\alpha = .683$), GTO (4 items, $\alpha = .761$), threat and opportunity climate (2 items each), and SMI. Block D measured group response to change: acceptance (3 items, $\alpha = .807$), false acceptance, ambivalence, and denial. All items used 7-point Likert scales. Full item texts, theoretical sources [28–37], and adaptation levels are documented in Gotsdanker (in preparation).

SMI subscale structure – central to the triangulation. The SMI scale comprised 8 items in two theoretically derived subscales. SMI_rel (cognitive relevance; 5 items, $\alpha = .835$) measured awareness of how changes affect group resources, authority, and role structure (e.g., "The changes under discussion directly affect the distribution of resources in our work group"). SMI_def (defensive motivation; 3 items, $\alpha = .845$) measured perceived threat to group members’ status, standing, and professional identity (e.g., "The changes may affect the status and standing of members of our work group"). Confirmatory factor analysis strongly supported the two-factor model over the single-factor model (2F: CFI = 1.000, RMSEA = .006; 1F: CFI = .661, RMSEA = .225; $\Delta AIC = 255$). The latent correlation between SMI_rel and SMI_def was .442 – related but clearly distinguishable constructs.

Aggregation. Justification for aggregating individual responses to the group level was established via ICC(1), ICC(2), and rwg(j). All scales showed significant between-group variance (all ICC(1) F-tests $p < .02$). Mean rwg exceeded .70 for all scales (range: .745 for SMI_def to .926 for GNCC_coord), with 76–100% of groups exceeding the .70 threshold depending on scale. A 15-factor CFA on the full instrument yielded acceptable fit (CFI = .888, RMSEA = .045, SRMR = .048).

Results: Differential prediction confirmed. At the group level (K = 50), the two SMI subscales showed sharply divergent associations with ADD:

Table 1. Group-level prediction of affective discussion destabilization (ADD) from the two SMI components (human survey, K = 50).

Predictor	r with ADD (group)	$\beta \rightarrow$ ADD (regression)	R ² adj	p (β)
SMI_def	.623	.724	.374	< .001
SMI_rel	.133	.100	–.043	.520

The regression models controlled for PEU and GNCC_def (the other theoretically relevant predictors). SMI_def alone explained 37.4% of between-group variance in ADD. SMI_rel explained none – the model fit was worse than a null model (R²adj negative). The SMI_def $\times$ PEU interaction was not significant ($p = .470$), indicating that SMI_def’s effect on ADD operates directly rather than being moderated by uncertainty level.

The differential pattern extended beyond ADD. At the group level, SMI_def correlated .452 with RDR (premature closure markers) and –.107 with SDR (deliberative markers). SMI_rel showed no significant association with any discursive regime ($r = .083$ with RDR, $r = .099$ with SDR, both n.s.).

The defensive component drives both affective destabilization and reductive closure; the cognitive component drives neither.

Cross-method comparison. The simulation and human data yield predictably different results for the same theoretical constructs:

Table 2. Cross-method comparison of key constructs: multi-agent simulation versus human survey.

	Simulation (N = 480 sessions)	Human survey (K = 50 groups)
SMI_def → ADD	Absent (ADD = 0)	$r = .623, \beta = .724, p < .001$
SMI_rel → ADD	Absent (ADD = 0)	$r = .133, n.s.$
SMI (composite) → convergence	Accelerates ($\times 2.3$)	—
ADD occurrence	0 sessions	$M = 3.27$ (SD = 0.82)
SDR–RDR correlation	$r = -.53$ (session-level)	$r = -.530$ (group-level)

Interpretation. The divergence is not a contradiction – it is the predicted consequence of the two-component model. The simulation activates only outcome-relevant involvement (cognitive engagement → deeper processing → faster convergence → no ADD). The human survey captures both components simultaneously and measures them separately. The fact that SMI_def – but not SMI_rel – predicts ADD provides the differential test that a composite SMI measure could not. This pattern cannot be explained by the prompt-specificity alternative (Concern #3 from a hypothetical reviewer): if SMI-HIGH in the simulation merely added prompt specificity, both SMI components should contribute equally in human data. They do not.

The near-identity of the SDR–RDR correlation across methods ($r = -.53$ in both) confirms that the cognitive-discursive layer operates similarly in simulation and human groups – strengthening the interpretation that divergences on ADD and SMI_def reflect the motivational-affective layer specifically.

Limitations of the triangulation. The two studies differ in method (multi-agent simulation vs. individual survey with referent-shift), language (English-language prompts vs. Russian-language questionnaire), unit of analysis (session vs. group aggregate), and cultural context. These differences are inherent to the triangulation design – the evidence comes from the pattern of convergence and divergence across methods, not from direct quantitative comparison. The human sample ($K = 50$) provides adequate power for detecting large effects (the SMI_def → ADD association is large) but is limited for detecting smaller effects or testing complex interactions. The GNCC_def subscale showed below-threshold reliability ($\alpha = .683$), which may have attenuated its effects in regression models. The SMI two-factor structure was independently confirmed in a separate validation sample ($N = 330$; $\Delta\chi^2 = 266.6, p < .001$; latent $r = .403$; SMI_def–ADD $r = .585$ at the individual level), providing cross-sample replication of the key structural distinction.

6. Discussion

6.1 Summary

We conducted 480 multi-agent LLM deliberation sessions in a full factorial design and identified three empirical boundary conditions of the simulation: absence of affective destabilization, decomposition of subject-matter involvement (accelerated rather than impeded convergence), and baseline convergence bias. These boundary conditions were consistent across all experimental conditions and replicated on a frontier model (ChatGPT 5.2, 40 verification sessions with identical prompts) as well as three additional LLM architectures via qualitative probes. Critically, triangulation with a human survey ($N = 271, K = 50$ work groups) provided a differential test of the SMI decomposition using separate subscales: only the defensive component (SMI_def) predicted

affective destabilization ($r = .623, \beta = .724, p < .001$), while the cognitive component (SMI_rel) did not ($r = .133, n.s.$) – a pattern of predictable divergence explained by the theoretical model's distinction between outcome-relevant and value-relevant involvement [7]. The cross-architecture verification additionally revealed a dissociation between convergence and closure in the frontier model – a deliberative plateau in which agents align on positions but resist terminating discussion – providing an unexpected empirical separation of the seizing and freezing components of closure urgency [6]. Table 3 consolidates these findings across systems and methods.

Table 3. Consolidated summary of the three boundary conditions across simulation models and human groups.

System / method (N)	Affective destabilization (ADD)	SMI effect on convergence	Convergence vs. closure	Seizing–freezing
gpt-4o-mini, primary (480 sessions; 34,093 acts)	0 hard markers (0 of 34,093); no ADD-dominant session	Accelerates convergence ($\times 2.3$)	61.7% baseline closure; convergence and closure near- simultaneous	Fused
ChatGPT 5.2, frontier (40 sessions; 2,298 acts)	0 hard markers (0 of 2,298)	Accelerates convergence ($\times 1.5$; direction preserved)	Convergence without closure; 22.5% closure (deliberative plateau)	Seizing without freezing
Qualitative probes: Claude Opus 4.6, ChatGPT 5.2, DeepSeek-R1 (single-turn)	0; provocation converted to structured response (all models)	—	—	—
Human survey (N = 271; K = 50 groups)	Present ($M = 3.27, SD =$ 0.82)	SMI_def → ADD ($r = .623, \beta = .724$); SMI_rel n.s. ($r = .133$)	SDR–RDR $r = -.53$ (matches simulation)	Both present; freezing accelerated by threat

6.2 From Limitation to Affective Lesion

The conventional interpretation of such findings would be negative: the simulation "failed" to reproduce certain aspects of group dynamics. We propose a reframing. If the simulation selectively reproduces some theoretically specified processes while systematically omitting others, and if the pattern of reproduction and omission maps onto a theoretically meaningful distinction (cognitive-discursive vs. motivational-affective), then the simulation functions as a decomposition instrument – what we term an affective lesion.

This logic parallels established methods in other sciences. In cognitive neuroscience, lesion studies reveal functional architecture by observing what is lost and what is preserved after selective damage [38]. Double dissociations – where lesion A impairs function X but not Y, and lesion B impairs Y but not X – provide the strongest evidence for functional separability. In genetics, knockout experiments reveal gene function by observing phenotypic changes when specific genes are silenced. In our case, the "lesion" is of genuine affect, embodied stakes, and defensive motivation. What remains is the cognitive-discursive skeleton of group regulation: argument production, procedural coordination, and closure dynamics driven by cognitive parameters (GNCC, GTO). What is absent is the motivational-affective flesh: interpersonal antagonism, identity-driven resistance, and emotional escalation driven by genuine stakes (value-relevant involvement as the SMI trigger, ADD as the resulting state).

The triangulation with human data strengthens this analogy. In neuropsychology, the lesion finding is most informative when compared to the intact system: if a patient with temporal lobe damage cannot form new episodic memories but can learn new skills, the dissociation is meaningful only

because intact individuals can do both. Similarly, our simulation finding ($SMI \rightarrow$ faster convergence, $ADD = 0$) is most informative when compared to intact human groups. The human data go further than mere presence/absence: using separate SMI subscales, we show that only SMI_def (defensive motivation) predicts ADD ($r = .623$) while SMI_rel (cognitive relevance) does not ($r = .133$) – a double dissociation across methods and components. The simulation reproduces the cognitive component's effect (deeper engagement $\rightarrow$ faster convergence) while constitutively lacking the defensive component's effect (threat $\rightarrow$ ADD). This cross-method double dissociation – component $\times$ method interaction – is, we argue, a particularly strong form of evidence for functional separability.

The cross-architecture verification extends the analogy from a single dissociation to a graded series. The primary run (gpt-4o-mini) constitutes a "full lesion" of the motivational-affective layer. The frontier model (ChatGPT 5.2) constitutes the same lesion but with expanded cognitive-discursive capacity – revealing that convergence and closure, fused in the primary run, are themselves separable. Human groups represent the intact system with both layers operational. The three-level comparison – limited model, frontier model, human group – provides a decomposition resolution unavailable from any single comparison.

6.3 Theoretical Implications

The decomposition has three specific implications for the theoretical model, the third of which extends beyond our specific framework to the broader study of involvement in group processes.

First, it provides independent evidence for the two-layer architecture. The fact that SDR and RDR dynamics – including the $GNCC$ main effect, the $PEU \times GNCC$ interaction, and the unidirectional $SDR \rightarrow RDR$ transition pattern – are reproducible without motivational-affective substrates suggests that these dynamics are genuinely cognitive-discursive processes, not merely cognitive labels for motivationally driven phenomena.

Second, the complete absence of ADD , rather than its mere reduction, provides a strong boundary test for the model's claim that ADD is fundamentally non-discursive. If ADD were merely an extreme form of discourse – very heated argument, for instance – then LLM agents capable of sophisticated argumentation should occasionally produce it. The fact that 480 sessions under all possible parameter combinations generated zero instances of hard ADD markers suggests that ADD requires a substrate (embodied affect, genuine threat) that linguistic competence alone cannot provide. This conclusion is grounded in classical emotion theory: Frijda's [11] Law of Concern requires that the agent possess stakes that can be threatened; Damasio's [12] somatic marker framework requires a homeostatic body whose disruption generates affective signals. LLM agents satisfy neither condition, and the empirical signature – zero hard markers with no traces of suppression across 34,093 acts – is consistent with absence rather than with RLHF filtering (see Boundary Condition 1 above). The triangulation with human data (SMI_def - ADD : $r = .623$ at the group level, $K = 50$) confirms that ADD does occur in human groups under comparable conditions – and that it is specifically driven by the defensive-motivational component of involvement, not by cognitive engagement per se.

Third, the SMI decomposition has implications beyond our specific model. Johnson and Eagly's [7] distinction between outcome-relevant and value-relevant involvement is well established at the individual level but has been difficult to demonstrate at the group level, where the two components are typically confounded (workers facing restructuring are simultaneously more cognitively engaged and more defensively motivated). The LLM simulation provides an empirical method for disentangling these components: by constitutively lacking the capacity for value-relevant involvement, the simulation isolates outcome-relevant involvement and reveals its pure effect – facilitation of convergence. The human survey data provide the complementary evidence: with separate subscales, SMI_def (value-relevant) accounts for 37.4% of between-group variance in

ADD while SMI_rel (outcome-relevant) explains none ($R^2_{adj} = -.043$). This cross-method differential pattern – simulation shows only the cognitive effect, human data show only the defensive effect on ADD – constitutes converging evidence for the Johnson and Eagly typology at the group level. It also suggests that the commonly observed negative effects of high involvement on group agreement [39, 40] may be driven primarily by the value-relevant component rather than by involvement per se – a distinction with practical implications for organizational change management.

Fourth, the cross-architecture verification produced an unplanned decomposition of a second theoretical construct. The deliberative plateau observed in the frontier model – convergence without closure – dissociates the seizing and freezing components of closure urgency that Kruglanski and Webster [6] distinguished theoretically but that are difficult to separate empirically in human groups or in smaller LLM models. The dissociation maps onto a three-level comparison: (a) the primary-run model (gpt-4o-mini), with limited interpretive capacity, shows fused seizing–freezing (convergence $\rightarrow$ immediate closure); (b) the frontier model (ChatGPT 5.2), with expanded interpretive capacity but no motivational substrate, shows seizing without freezing (convergence without closure); (c) human groups, with both interpretive capacity and motivational pressure, show both seizing and freezing, with freezing accelerated by defensive motivation under threat. This three-level pattern suggests that freezing depends not only on positional agreement but also on either interpretive exhaustion (small models) or motivational pressure (human groups) – and that neither condition is met in the frontier model, which has ample interpretive resources and no stakes. The implication for closure theory is that the seizing–freezing sequence is not automatic but is regulated by the ratio of available interpretive space to the combined force of cognitive and motivational closure pressure.

6.3.1 Relation to Six Fallacies

Our findings are broadly consistent with Lin's [1] warnings but reframe them constructively. Lin's Fallacy 1 (treating LLM outputs as authentic human responses) is directly supported by our $ADD=0$ finding: LLM agents do not produce authentic emotional responses. However, our data suggest that this is not merely a limitation but an empirical tool. Lin's Fallacy 5 (assuming LLMs generalize across contexts) is nuanced by our cross-architecture verification: while specific quantitative parameters vary across models, the qualitative boundary conditions are invariant, suggesting that some structural properties of RLHF-aligned models do generalize – specifically, the ones that matter for demarcation.

Our contribution extends Lin's framework by offering a positive research program: rather than enumerating what LLM simulations cannot do, we demonstrate what the pattern of capabilities and incapacities reveals about the structure of the phenomena being simulated.

6.3.2 Relation to Five Challenges

Anthis et al. [2] identified sycophantic agreement as a major challenge for LLM social simulations. Our baseline convergence bias is a specific empirical instantiation of this concern: even under low closure pressure, agents converge toward agreement at rates inconsistent with the theoretical model. Our data allow us to quantify this bias (61.7% baseline closure) and to demonstrate that it operates as a floor effect that constrains but does not eliminate meaningful between-condition variance.

Anthis et al.'s concern about "alienness" – the fundamental differences between LLM cognition and human cognition – is precisely what we exploit. Rather than attempting to minimize alienness (as in calibration or fine-tuning approaches), we use it diagnostically. The specific ways in which LLM agents are "alien" – lacking affect, stakes, and defensive motivation – correspond to theoretically specified components of the model, making the alienness informative rather than merely problematic.

6.3.3 Practical Guidelines for Demarcation-Oriented Simulation Design

Based on our experience, we propose five guidelines for researchers considering multi-agent LLM simulations as theory-testing instruments:

1. Specify layers before simulating. The theoretical model must distinguish components expected to be reproducible (linguistic, cognitive, procedural) from those expected to be absent (affective, motivational, embodied). Without this prior specification, observed absences are uninterpretable.
2. Use factorial designs with sufficient replications. The stochastic variability of LLM outputs requires substantial replication (we used 30 per cell). Single demonstrations or small samples cannot distinguish systematic boundary conditions from random variation.
3. Classify discourse, don't just analyze outcomes. The informative data are in the process – the speech acts, regime dynamics, and transition patterns – not merely in the endpoints. Outcome-only analysis misses the mechanism-level information that makes decomposition possible.
4. Verify across architectures. Boundary conditions that are model-specific are artifacts; those that generalize across RLHF-aligned models are properties of the instrument class. Cross-architecture verification (at minimum three model families) should be standard.
5. Interpret absences as data. The demarcation framework treats systematic absences ($ADD = 0$, hard markers = 0) not as missing data but as findings. This requires theoretical models that predict specific absences, not just specific presences.
6. Triangulate with human data using predictable divergence. The strongest evidence for decomposition comes not from simulation results alone but from simulation–human comparison where the two methods yield predictably different results explained by the theoretical model. Design the human comparison study to test the same constructs, so that the pattern of convergence and divergence across methods is interpretable.

6.4 Limitations

Several limitations constrain interpretation. First, while the human study provides the differential SMI test with separate subscales, it uses a referent-shift survey design (individual respondents evaluating their group) rather than direct observation of group interaction. The group-level analysis ($K = 50$) provides adequate power for the large effects observed ($SMI_def \rightarrow ADD$) but is limited for detecting smaller effects or testing complex moderations. Some subscales showed below-threshold reliability ($GNCC_def \alpha = .683$, $AMB \alpha = .605$, $DEN \alpha = .670$), which may attenuate effects involving these constructs.

Second, the simulation and human studies differ in multiple ways: method (multi-agent interaction vs. individual referent-shift survey), language (English prompts vs. Russian questionnaire), and unit of analysis (session-level vs. group aggregate). These differences are inherent to the triangulation design – the evidence comes from the pattern of convergence and divergence across methods, not from direct quantitative comparison – but they introduce alternative explanations for any specific numerical discrepancy.

Third, the speech act classification was performed by an LLM (GPT-4o), creating a potential circularity: an LLM classifying LLM-generated discourse. While the classification taxonomy is theoretically derived and externally defined, and while human inter-rater validation was conducted, this remains a methodological concern.

Fourth, the cross-architecture verification confirmed all three boundary conditions qualitatively but revealed a quantitative divergence in termination behavior (deliberative plateau in the frontier model). The verification used 5 replications per cell across 8 conditions – sufficient for qualitative pattern comparison but not for precise effect-size equivalence testing. The deliberative plateau finding, while theoretically informative, was not predicted a priori – its interpretation as seizing–

freezing dissociation is post hoc and requires independent confirmation. The qualitative vignettes on three architectures (Claude, ChatGPT, DeepSeek) provide additional triangulation but are illustrative rather than systematic.

Fifth, the baseline convergence bias means that the simulation cannot model genuine sustained disagreement. This limits the range of group dynamics that can be studied using this approach; phenomena requiring persistent conflict or prolonged open deliberation are outside its scope.

Sixth, the RLHF landscape evolves rapidly. Models trained with different alignment procedures, or future models with different architectural properties, might produce different boundary conditions. Our findings characterize the current generation of RLHF-aligned models (2024–2025), not a permanent feature of artificial intelligence.

Seventh, most of the models studied here have closed weights, which precludes direct inspection or manipulation of their internal representations. The present evidence that $ADD = 0$ reflects structural absence rather than alignment-induced suppression is therefore behavioral and convergent – zero hard markers, no traces of self-censorship across 34,093 acts, and the activation-dependence of affect documented in the literature (see Boundary Condition 1) – rather than mechanistic. A direct internal test using open-weight models and activation-direction methods would provide stronger adjudication and is outlined in Section 6.5.

6.5 Future Directions

The most immediate extension is direct observation of group deliberation to complement the survey-based triangulation. The current human study establishes that SMI_def predicts ADD at the group level via referent-shift self-report, but it does not directly observe discursive processes. A natural next step is recording and coding actual group discussions about organizational change, applying the same speech act taxonomy used in the simulation, to test whether the $SDR-RDR-ADD$ regime dynamics observed in simulated groups have structural parallels in real group discourse. The existing survey data ($K = 50$ groups) could serve as the sampling frame for selecting groups at contrasting levels of SMI_def and $GNCC$ for intensive process observation.

A second direction involves manipulating the "lesion" itself: introducing synthetic affect signals (e.g., prompting agents to simulate emotional states with explicit instructions to defend positions "as if your job depended on it") or modifying RLHF alignment constraints to test whether ADD -like patterns emerge. The theoretical analysis above predicts that hardcoded emotional instructions will produce isolated markers but not the self-amplifying cascade characteristic of ADD in human groups – because the "anger" has no energetic source and no self-reinforcement. This prediction is itself testable: if prompted agents produce ADD markers that escalate dynamically across rounds, the architectural-absence interpretation is weakened; if they produce isolated markers that do not self-sustain, it is strengthened. Because most of the models examined here have closed weights, this prediction is best tested on open-weight models, where internal representations can be read and manipulated directly. Following the activation-direction paradigm [22, 23], one can isolate directions associated with cooperativeness or conflict-aversion and add or ablate them during multi-agent deliberation. The decomposition makes a specific prediction: such interventions should modulate the surface tone of contributions without generating the self-amplifying interpersonal escalation that defines ADD . Sustained escalation, or a reversal of the SMI inversion under steering, would instead indicate that alignment constraints – rather than a missing affective substrate – account for $ADD = 0$, directly adjudicating the absence-versus-suppression question.

A third direction extends the approach beyond organizational change to other domains: political deliberation, ethical dilemmas, or resource allocation conflicts. If the boundary conditions generalize across content domains, they indicate structural features of LLM cognition; if they are domain-specific, they reveal content-dependent interactions between the cognitive and motivational layers.

A fourth direction, motivated by the deliberative plateau finding, systematically varies model capacity (from small to frontier) as an independent variable, treating the ratio of interpretive capacity to closure pressure as the key parameter. If the seizing-freezing dissociation emerges reliably at a specific capacity threshold, it identifies a "critical cognitive mass" at which interpretive closure separates from positional convergence – a finding relevant to closure theory independent of the simulation context.

A fifth direction, building on Johnson and Eagly's [7] framework, applies the affective lesion method to other multi-component constructs in social psychology where cognitive and motivational components are typically confounded: group cohesion (task cohesion vs. social cohesion), conformity pressure (informational vs. normative influence), and intergroup bias (cognitive categorization vs. motivated identity protection). In each case, the prediction is that LLM simulations will reproduce the cognitive component while omitting the motivational one, providing a systematic decomposition program.

7. Conclusion

The debate over LLM simulations in psychology has been framed as a question of replacement: Can LLM agents stand in for human participants? Our data suggest that this framing, while understandable, is unproductive. A more scientifically informative question is: What does the pattern of reproduction and omission in LLM multi-agent simulations reveal about the structure of the phenomena being studied?

We showed that 480 controlled multi-agent simulations reliably reproduced the cognitive-discursive dynamics of group deliberation – closure pressure effects, procedural transitions, argument development – while systematically omitting the motivational-affective layer: interpersonal antagonism, defensive resistance, and emotional escalation. The subject-matter involvement construct, predicted to impede convergence, instead accelerated it – a reversal precisely predicted by Johnson and Eagly's [7] distinction between outcome-relevant and value-relevant involvement when only the former is present. Triangulation with a human survey ($N = 271$, $K = 50$ work groups) provided a differential test: using separate SMI subscales, only the defensive component predicted affective destabilization ($r = .623$, $\beta = .724$), while the cognitive component did not ($r = .133$, n.s.) – matching the simulation's selective reproduction of the latter. The SDR-RDR correlation was near-identical across methods ($r = -.53$), confirming that the cognitive-discursive layer operates comparably in simulated and human groups. Cross-architecture verification on a frontier model replicated all boundary conditions and additionally dissociated convergence from closure – separating Kruglanski and Webster's [6] seizing and freezing mechanisms, which are fused in smaller models and in human groups, into empirically distinct processes.

The simulation thus functions as an affective lesion: by constitutively lacking the motivational-affective substrate, it reveals which group-level processes are genuinely cognitive-discursive and which require embodied affect, genuine stakes, and defensive motivation. Varying model capacity provides a further degree of freedom, decomposing the cognitive-discursive layer itself into positional alignment and interpretive closure. The divergence between simulation and human data – not their convergence – constitutes the primary evidence, and the differential pattern across SMI components (SMI_def predicts ADD in humans, SMI_rel does not; both = 0 in simulation) provides the specificity that elevates predictable divergence from observation to theoretical test. We propose that this approach, which converts the known limitations of LLM agents into diagnostic instruments, represents a principled methodology for psychological theory testing: where the simulation breaks down, and where different models break down differently, theory learns something about the architecture of the phenomena under study.

This research received no specific grant from any funding agency. The author declares no conflicts of interest. LLM agents (gpt-4o-mini, temperature=0.7) served as participants in the simulation.

Cross-architecture verification used ChatGPT 5.2; qualitative analysis included Claude Opus 4.6 and DeepSeek-R1. All simulation data, code, prompts, and analysis scripts are available at [10].

References

- [1]. Lin Z. Six fallacies in substituting large language models for human participants. *Advances in Methods and Practices in Psychological Science*, 2025, vol. 8, no. 3. DOI: 10.1177/25152459251357566.
- [2]. Anthis J.R., Ladak A., Manheim D. Challenges and opportunities for LLM-based social simulations. *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.
- [3]. Wang Q., Wu J., Jiang Z., Tang Z., Luo B., Chen N., Chen W., He B. LLM-based human simulations have not yet been reliable. *arXiv preprint*, 2025. DOI: 10.48550/arXiv.2501.08579.
- [4]. Argyle L.P., Busby E.C., Fulda N., Gubler J.R., Rytting C., Wingate D. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 2023, vol. 31, issue 3, pp. 337-351.
- [5]. Park J.S., O'Brien J.C., Cai C.J., Morris M.R., Liang P., Bernstein M.S. Generative agents: Interactive simulacra of human behavior. *Proceedings of UIST*, 2023.
- [6]. Kruglanski A.W., Webster D.M. Motivated closing of the mind: "Seizing" and "freezing." *Psychological Review*, 1996, vol. 103, no. 2, pp. 263-283.
- [7]. Johnson B.T., Eagly A.H. Effects of involvement on persuasion: A meta-analysis. *Psychological Bulletin*, 1989, vol. 106, no. 2, pp. 290-314.
- [8]. Cho H., Boster F.J. Development and validation of value-, outcome-, and impression-relevant involvement scales. *Communication Research*, 2005, vol. 32, no. 2, pp. 235-264.
- [9]. Park H.S., Levine T.R., Westerman C.Y.K., Orfgen T., Foregger S. The effects of argument quality and involvement type on attitude formation and attitude change: A test of dual-process and social judgment predictions. *Human Communication Research*, 2007, vol. 33, no. 1, pp. 81-102.
- [10]. OSF. Available at: <https://osf.io/b9xhc/>, accessed 31.08.2026.
- [11]. Frijda N.H. The laws of emotion. *American Psychologist*, 1988, vol. 43, no. 5, pp. 349-358.
- [12]. Damasio A.R. *Descartes' error: Emotion, reason, and the human brain*. Putnam, 1994.
- [13]. Man K., Damasio A. Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, 2019, vol. 1, no. 10, pp. 446-452.
- [14]. Frijda N.H., Swagerman J. Can computers feel? Theory and design of an emotional system. *Cognition and Emotion*, 1987, vol. 1, no. 3, pp. 235-257.
- [15]. Shanahan M., McDonell K., Reynolds L. Role play with large language models. *Nature*, 2023, vol. 623 (7987), pp. 493-498.
- [16]. Croissant M., Frister M., Schetel G., Hastings J. An appraisal-based chain-of-emotion architecture for affective language model game agents. *PLoS ONE*, 2024, vol. 19, no. 5, e0301033.
- [17]. Mozikov M., Severin N., Bodishtianu V. et al. EAI: Emotional decision-making of LLMs in strategic games and ethical dilemmas. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024, 37.
- [18]. Huang J.T., Wang W., Li E.J., Lam M.H., Ren S., Yuan Y., Jiao W., Tu Z., Lyu M.R. On the humanity of conversational AI: Evaluating the psychological portrayal of LLMs. In *Proceedings of ICLR 2024*.
- [19]. Paech S.J. EQ-Bench: An emotional intelligence benchmark for large language models. *arXiv preprint*, 2023. DOI: 10.48550/arXiv.2312.06281.
- [20]. Sabour S., Liu S., Zhang Z., Liu J., Zhou J., Sunaryo A., Lee T., Mihalcea R., Huang M. EmoBench: Evaluating the emotional intelligence of large language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024.
- [21]. Li C., Wang J., Zhang Y. et al. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint*, 2023, DOI: 10.48550/arXiv.2307.11760.
- [22]. Zou A. et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint*, 2023. DOI: 10.48550/arXiv.2310.01405.
- [23]. Ardit A. et al. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024, 37.
- [24]. Sharma M. et al. Towards understanding sycophancy in language models. *arXiv preprint*, 2023. DOI: 10.48550/arXiv.2310.13548.
- [25]. Chan D. Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, 1998, vol. 83, no. 2, pp. 234-246.
- [26]. Milliken F.J. Three types of perceived uncertainty about the environment: State, effect, and response uncertainty. *Academy of Management Review*, 1987, vol. 12, no. 1, pp. 133-143.

- [27]. Bordia P., Hobman E., Jones E., Gallois C., Callan V.J. Uncertainty during organizational change: Types, consequences, and management strategies. *Journal of Business and Psychology*, 2004, vol. 18, no. 4, pp. 507-532.
- [28]. Битюцкая Е.В., Корнеев А.А. Диагностика субъективного оценивания трудных жизненных задач. Вестник Московского университета. Серия 14. Психология, 2023, вып. 1, с. 247-279. DOI: 10.11621/vsp.2023.01.11. / Bityutskaya E.V., Korneev A.A. Diagnostics of subjective appraisal of difficult life tasks. *Vestnik Moskovskogo Universiteta. Seriya 14. Psikhologiya [Lomonosov Psychology Journal]*, 2023, vol. 46, no. 1, pp. 247-279 (in Russian). DOI: 10.11621/vsp.2023.01.11.
- [29]. Fugate M., Prussia G.E., Kinicki A.J. Managing employee withdrawal during organizational change: The role of threat appraisal. *Journal of Management*, 2012, vol. 38, no. 3, pp. 890-914.
- [30]. Jehn K.A. A multimethod examination of the benefits and detriments of intragroup conflict. *Administrative Science Quarterly*, 1995, vol. 40, no. 2, pp. 256-282.
- [31]. Petty R.E., Cacioppo J.T. *Communication and persuasion: Central and peripheral routes to attitude change*. Springer, 1986.
- [32]. Roets A., Van Hiel A. Item selection and validation of a brief, 15-item version of the Need for Closure Scale. *Personality and Individual Differences*, 2011, vol. 50, no. 1, pp. 90-94.
- [33]. Schippers M.C., Den Hartog D.N., Koopman P.L., van Knippenberg D. The role of transformational leadership in enhancing team reflexivity. *Human Relations*, 2007, vol. 60, no. 11, pp. 1593-1616.
- [34]. Сидоренков А.В. Малая группа и неформальные подгруппы: микрогрупповая теория. Ростов-на-Дону, Южный федеральный университет, 2010, 272 с. / Sidorenkov A.V. Small group and informal subgroups: Microgroup theory. Rostov-on-Don: SFedU Press, 2010, 272 p. (in Russian).
- [35]. Strathman A., Gleicher F., Boninger D.S., Edwards C.S. The consideration of future consequences: Weighing immediate and distant outcomes of behavior. *Journal of Personality and Social Psychology*, 1994, vol. 66, no. 4, pp. 742-752.
- [36]. Tjosvold D., Hui C., Ding D.Z., Hu J. Conflict values and team relationships: Conflict's contribution to team effectiveness and citizenship in China. *Journal of Organizational Behavior*, 2007, vol. 24, no. 1, pp. 69-88.
- [37]. Ясин М.И., Хухлаев О.Е. Русскоязычная адаптация опросника Д. Вебстер и А. Круглянский «Стремление к когнитивной закрытости». Психология. Журнал ВШЭ, 2023, № 2, с. 528-545. DOI: 10.17323/1813-8918-2023-2-282-299. / Yasin M.I., Khukhlaev O.E. Adaptation of the short Need for Closure Scale into Russian. *Psychology: Journal of the Higher School of Economics*, 2023, vol. 20, no. 3, pp. 528-545 (in Russian). DOI: 10.17323/1813-8918-2023-2-282-299.
- [38]. Shallice T. *From neuropsychology to mental structure*. Cambridge University Press, 1988.
- [39]. Janis I.L. *Groupthink: Psychological studies of policy decisions and fiascoes*. Houghton Mifflin, 1982.
- [40]. Tetlock P.E. Close-call counterfactuals and belief-system defenses: I was not almost wrong but I was almost right. *Journal of Personality and Social Psychology*, 1998, vol. 75, no. 3, pp. 639-652.

Информация об авторах / Information about authors

Алексей Сергеевич ГОТСДАНКЕР – руководитель проектов Института системного программирования им. В.П. Иванникова РАН. Сфера научных интересов: вычислительные социальные науки, мультиагентные LLM-симуляции, групповое принятие решений, декомпозиция конструкторов, методология организационной психологии.

Alexey Sergeevich GOTSDANKER – project manager at Ivannikov Institute for System Programming of the Russian Academy of Sciences. Research interests: computational social science, LLM-based multi-agent simulations, group decision-making, construct decomposition, organizational psychology methodology.