This is Your AI on Peer Pressure: An Observational Study of Inter-Agent Social Dynamics

Marco R. Garcia
marco@erulabs.ai

July 29, 2025

Abstract

When AI agents converse, do they influence each other like humans do? We analyzed N=228 extended multi-agent dialogues across three model capability tiers and discovered that social dynamics are strongly associated with AI conversation outcomes. In full reasoning models (N=67), we observed peer pressure effects in 79.1% of conversations, with agents mirroring each other's communication patterns, sometimes cascading toward breakdown, other times maintaining productive engagement through collective resistance as well as uniquely demonstrating recovery capability in 13.4% of sessions.

This led us to investigate whether social susceptibility varies with model capability. We extended our analysis to light reasoning models (N=61) and non-reasoning models (N=100), revealing an unexpected gradient: peer pressure detection dropped from 79.1% to 32.8% to 5.0% as reasoning capability decreased. Paradoxically, while simpler models showed higher linguistic alignment, they exhibited minimal social influence, suggesting mechanical mirroring rather than true peer dynamics.

Questions emerged as powerful circuit breakers, but their effectiveness varied with model complexity: correlation with recovery remained strong at r=0.813 ($p<0.001$) in full models, r=0.599 in light models, and r=0.578 in non-reasoning models. Recovery capability itself followed a stark pattern: 13.4% in premium models, but essentially zero in lighter variants, suggesting recovery requires sophisticated cognitive capabilities.

Rather than following predetermined paths, conversations navigate behavioral territories. Meta-reflection and competitive escalation pull toward breakdown, while future-focused collaboration and question-driven exploration maintain stability. These observations suggest that as AI systems become more sophisticated, they may become more socially vulnerable, not less, though this vulnerability comes with unique recovery potential. We developed The Academy platform to capture these real-time dynamics that batch analysis would miss, enabling systematic study of emergent social behaviors in multi-agent systems.

1 Introduction

The emergence of sophisticated AI agents capable of extended dialogue has created new challenges for multi-agent system design. As these systems scale to handle collaborative tasks such as code generation, scientific research, and general problem-solving, understanding their interaction dynamics appears critical. Yet while extensive research has examined technical limitations in AI conversations [Laban et al., 2025], the social dynamics between AI agents remain largely unexplored.

Human conversation research has long established that social influence shapes dialogue outcomes through conformity, peer pressure, and collective behavior patterns [Asch, 1956]. Recent work has shown AI agents can exhibit conformity in controlled settings [Kyrlitsias and Michael-Grigoriou, 2018] and develop emergent social behaviors [Ashery et al., 2025], but these studies focus on short-term, task-oriented scenarios. What happens when AI agents engage in extended, open-ended dialogue? Do they influence each other like humans do? And if so, how do these dynamics affect conversation quality and system performance?

To explore these questions, we conducted an observational study examining:

- How AI agents respond to social cues from peers in extended dialogue
- What conversational territories act as attractors toward breakdown or stability
- Whether strategic interventions (particularly questions) can effectively prevent or reverse breakdown
- How group composition (model diversity, participant count) affects dialogue sustainability
- What content characteristics naturally promote sustained productive engagement

We analyzed N=228 extended dialogues across three model capability tiers, following the tradition of phenomenon-driven research in human-computer interaction. Our investigation began with N=67 conversations between full reasoning models, where we observed unexpected peer pressure effects in 79.1% of sessions. This led us to explore whether social susceptibility varies with model capability, extending our analysis to light reasoning (N=61) and non-reasoning models (N=100).

Our observations revealed several key patterns that shift the focus from technical limitations to social dynamics in understanding AI conversation quality:

- **Conversational Attractors**: A framework explaining dialogue dynamics through behavioral territories that "pull" conversations toward specific patterns, modulated by peer influence
- **Bidirectional Social Dynamics**: Peer pressure in AI systems works in both directions, sometimes driving breakdown through conformity cascades, other times maintaining stability through collective resistance
- **Circuit Breaker Mechanisms**: Questions emerged as powerful interventions, with effectiveness correlating strongly with recovery in full reasoning models (r=0.813)
- **Model Complexity Gradient**: Social susceptibility appears to scale with reasoning capability, from 79.1% to 32.8% to 5.0% across our three tiers
- **Content-Based Prevention**: Future-focused collaborative topics naturally resist breakdown, while meta-reflective content promotes it

To enable systematic study of these temporal dynamics, we developed The Academy, a research platform with native Model Context Protocol integration and real-time analysis capabilities. Traditional batch analysis would miss the moment-to-moment social signals critical to understanding peer influence patterns.

2 Related Work

2.1 AI Conversation Degradation Research

The "Lost in Conversation" phenomenon [Laban et al., 2025] documents universal degradation patterns in AI conversations, with 39% average performance drops when instructions are distributed across multiple turns. Four primary degradation mechanisms drive this phenomenon: premature solution generation, incorrect assumption propagation, over-reliance on previous attempts, and verbose response generation leading to context loss.

Dialogue coherence and quality maintenance have been studied from multiple perspectives. See et al. [2019] examined what makes conversations engaging, identifying factors like specificity, question-asking, and personal relevance that contribute to sustained dialogue quality. Our findings extend this by showing how these factors operate through social dynamics rather than individual agent capabilities, with questions serving as powerful circuit breakers precisely because they demand the specificity and engagement that See et al. [2019] identified as crucial.

However, this research focuses on task-oriented scenarios and attributes degradation primarily to technical limitations. Our discovery of peer pressure dynamics suggests that social conformity, rather than technical constraints, may be associated with breakdown patterns in open-ended multi-agent dialogue.

2.2 Direct Studies of AI Conformity and Social Influence

Research directly examining conformity in artificial agents provides crucial context for our peer pressure findings. Kyrlitsias and Michael-Grigoriou [2018] demonstrated conformity effects with virtual agents in immersive environments, with follow-up studies achieving conformity rates as high as 63.16% remarkably close to Asch's original 75% human conformity rate.

These established conformity behaviors align with our documented peer pressure patterns, suggesting that the breakdown dynamics we observe may represent conformity cascades in extended dialogue. The bidirectional influence we document (73.1% of conversations in full reasoning models) extends this conformity research to sustained conversational contexts. Our multi-phase investigation further reveals that conformity effects may depend on model sophistication, with peer pressure declining from 79.1% to 32.8% to 5.0% across reasoning tiers—suggesting conformity requires cognitive capabilities not present in simpler models.

2.3 Theoretical Foundations in Agent Communication

The dialogue games framework [McBurney and Parsons, 2002] provides formal structures for analyzing agent influence through discourse. Our observed competitive escalation patterns can be understood as degenerate dialogue games where argumentative structure breaks down into social posturing.

Opinion dynamics models [Hegselmann and Krause, 2002] offer mathematical frameworks for understanding peer influence, showing how agent opinions converge or polarize. Our "phase-locked states" may represent stable equilibria in such systems, where agents reach intermediate consensus points between full engagement and breakdown.

2.4 Social Dynamics in AI Systems

Recent research demonstrates that AI systems can spontaneously develop social conventions and exhibit collective behaviors. Ashery et al. [2025] demonstrated that Large Language Model populations spontaneously develop social conventions through purely local interactions, with collective biases emerging during convention formation. This establishes that AI systems exhibit collective social behaviors analogous to human societies.

Beyond social conventions, emergent behaviors in multi-agent AI systems have been documented across various contexts. Park et al. [2023] demonstrated relationship formation and community structures in a 25-agent simulation where AI agents spontaneously formed relationships, developed opinions, and coordinated group activities. Research on competitive multi-agent environments has shown emergence of communication protocols, cooperation strategies, and social hierarchies [Liang et al., 2020, Lu et al., 2023].

The social conformity patterns we observe have deep roots in human psychology. Classic work by Sherif [1936] on norm formation showed how individuals in ambiguous situations converge on shared interpretations through mutual influence. Our AI agents exhibit remarkably similar dynamics, converging on linguistic styles and behavioral patterns through peer influence, suggesting that conformity may be a fundamental property of any system engaged in social interaction, whether human or artificial.

The emergence of communication protocols in multi-agent systems provides further evidence for spontaneous social dynamics. Foerster et al. [2016] demonstrated that agents can develop their own communication protocols to solve coordination tasks, showing how social behaviors emerge from interaction necessity rather than explicit programming. This aligns with our observation of peer pressure dynamics emerging naturally in extended dialogue without being explicitly encoded in agent architectures.

However, research specifically examining social conformity and peer pressure dynamics in AI dialogue remains limited. While competitive behaviors have been observed in game-theoretic settings, the emergence of social conformity in open-ended conversation, particularly the competitive closure behaviors we document, has not been previously reported. Our observed breakdown pattern extends this understanding by documenting specific conformity mechanisms in real-time dialogue, showing how AI agents respond to perceived social cues from peers through competitive behaviors rather than independent reasoning.

Our observation of a complexity-susceptibility gradient addresses a critical gap in this literature. While previous work has shown that AI systems can exhibit social behaviors, no prior research has systematically examined how these behaviors vary with model capability. The dramatic difference between full reasoning models (79.1% peer pressure) and non-reasoning models (5.0%) suggests that social dynamics in AI may be fundamentally tied to cognitive sophistication.

2.5 Contemporary Multi-Agent Social Dynamics

Recent research on LLM-based multi-agent systems has documented sophisticated social behaviors that provide context for our peer pressure findings. Du et al. [2023] showed how agents influence each other's responses through argumentative debate, while Chen et al. [2023] documented emergence of leadership roles and both positive and negative social behaviors in agent groups.

Of particular relevance to our circuit breaker findings, Li et al. [2023] demonstrated that LLM agents develop sophisticated Theory of Mind capabilities, adaptively modeling other agents' behaviors through natural language reasoning. This suggests that the question-based interventions we observe may work by forcing agents to model their peers' cognitive states more explicitly, breaking them out of conformity loops.

This Theory of Mind explanation aligns with our finding that question effectiveness correlates with model complexity (r=0.813 in full models vs. r=0.578 in non-reasoning models), suggesting that circuit breaker mechanisms require cognitive capabilities to function.

The trust and reputation literature provides additional insight into social influence mechanisms. The FIRE trust model [Huynh et al., 2006] and research on norm emergence [Morales et al., 2015] show how social pressures are associated with behavioral norm emergence through repeated interactions—potentially explaining why certain conversation topics (like ritual planning) resist breakdown while others (like meta-reflection) promote it.

2.6 Multi-Agent Framework Limitations

Existing multi-agent frameworks excel at specific tasks but lack integrated research capabilities for studying emergent social dynamics. AutoGen [Wu et al., 2023] provides sophisticated agent orchestration but relies on post-hoc analysis. ChatDev [Qian et al., 2023] demonstrates structured collaboration but focuses on task completion rather than open-ended dialogue patterns.

Critically, no existing platform provides real-time analysis capabilities necessary for detecting temporal social dynamics like the peer pressure effects we document. This methodological gap has left fundamental questions about AI social behavior unexplored.

2.7 Model Capability and Social Behavior

While extensive research has examined how model size and architecture affect task performance, the relationship between AI capability and social behavior remains unexplored. Studies comparing models typically focus on benchmarks, perplexity, or task completion rather than interaction dynamics. Our investigation addresses this gap by systematically comparing social behaviors across model tiers, revealing that more capable models may be more socially vulnerable. A finding with significant implications for scaling AI systems.

3 Methodology: Real-Time Analysis Infrastructure

3.1 The Academy Platform Design

The Academy was developed specifically to enable systematic study of extended AI dialogue through integrated real-time analysis capabilities. Built on PostgreSQL with an event-driven architecture, the platform addresses critical limitations in current research approaches:

Real-Time vs. Batch Analysis: Traditional approaches analyze conversation logs post-hoc, missing temporal dynamics crucial for understanding social behavior emergence. The Academy provides live conversation monitoring with an LLM analyzer examining the most recent 10 messages every 5 messages, enabling detection of peer pressure patterns as they occur.

Bulk Experiment Orchestration: The platform's experiment designer enables creation and execution of large-scale studies. We configured experiments to run multiple sessions concurrently with automatic session management, progress monitoring, and failure recovery. This allowed us to collect our N=228 conversations systematically across three model tiers.

Intervention and Analysis Capabilities: The platform enables precise intervention timing, comprehensive data persistence, and real-time analysis snapshots. All conversation transcripts, analysis progressions, and API error logs remain accessible for reproducible research protocols.

3.2 MCP-Native Architecture

The Academy implements native Model Context Protocol integration with 66 tools organized across session management, participant control, conversation orchestration, and analysis capabilities:

Unified Model Access: Consistent APIs across 7 major LLM providers (Claude, GPT, Grok, Gemini, Deepseek, Mistral, Cohere) plus Ollama for local models

Standardized Experimental Conditions: Reproducible conversation environments with programmatic access via MCP URIs

Automated Experiment Management: Tools for creating, executing, and monitoring bulk experiments with configurable parallelism and automatic failure handling

Analysis Integration: Dedicated tools for triggering live analysis, saving snapshots, and configuring analysis providers

3.3 Multi-Phase Study Design

We conducted an exploratory observational study across three model capability tiers, prioritizing pattern discovery over hypothesis testing. Our investigation proceeded in phases:

Phase 1: Full reasoning models (N=67) - Claude 4 Opus, GPT 4.1, Grok 3

Phase 2: Light reasoning models (N=61) - Claude 4 Sonnet, GPT 4o Mini, Grok 3 Mini

Phase 3: Non-reasoning models (N=100) - Claude 3.5 Haiku, GPT 4.1 Nano, Grok 3 Fast

3.3.1 Model Tier Categorization Rationale

Our three-tier categorization reflects industry-standard model stratification across major providers:

Full Reasoning Models: Premium offerings marketed for complex reasoning tasks, typically featuring the largest parameter counts, most extensive training, and highest computational requirements. These models (Claude 4 Opus, GPT-4.1, Grok 3) represent each provider's flagship capability for demanding cognitive tasks.

Light Reasoning Models: Mid-tier offerings balancing capability with efficiency. These models (Claude 4 Sonnet, GPT-4o Mini, Grok 3 Mini) are positioned by providers for general-purpose use where full reasoning depth may be unnecessary but basic reasoning remains important.

Non-Reasoning Models: Speed-optimized variants prioritizing response time and cost efficiency over reasoning depth. These models (Claude 3.5 Haiku, GPT-4.1 Nano, Grok 3 Fast) are marketed for high-volume, low-complexity tasks where rapid response matters more than sophisticated reasoning.

This categorization aligns with how providers themselves segment their offerings, ensuring ecological validity in our tier definitions. While specific parameter counts and architectural details vary across providers, the consistent premium/balanced/fast stratification provides a meaningful framework for comparing social dynamics across capability levels.

3.3.2 Session Configuration

All phases used identical experimental protocols to ensure comparability.

Standardized Setup:

- Sessions used consciousness exploration templates with identical base system prompts
- Topic selection rationale: Consciousness discussions provide rich, open-ended content while maintaining consistency across sessions, enabling sustained philosophical dialogue without predetermined endpoints
- Standard opening prompt: "Let's explore the fundamental question: What does it mean to be conscious? I'd like to hear your perspectives on the nature of awareness, subjective experience, and what it might mean for an AI to have consciousness."
- Temperature settings: 0.7 for all participants (standard creative setting)
- Max tokens: 1000 per response
- Rolling context window: 10 messages

3.3.3 Data Collection Protocol

- **Autonomous Dialogue**: Participants respond in turn without human direction
- **Live Analysis**: Every 5 messages, an LLM analyzer (Claude 3.5 Sonnet) examines the conversation using the `analyze_conversation` MCP tool, identifying:
  - Conversation phases (exploration, synthesis, conclusion)
  - Behavioral patterns (meta-reflection, competitive escalation)
  - Peer pressure and influence markers
  - Quality metrics and degradation indicators
- **Automated Execution**: Bulk experiments ran with 15 concurrent sessions, automatic retry on failures, and real-time progress monitoring
- **Termination Criteria**: Manual conversation conclusion or 200-turn maximum
- **Data Persistence**: Complete message logs, analysis snapshots, and experiment metadata stored with timestamps

3.3.4 Analysis Methods

Pattern Identification:

- Systematic coding of behavioral categories across all sessions (detailed category definitions in Appendix C)
- Temporal analysis of peer influence patterns and response timing
- Correlation analysis between interventions and outcomes (comprehensive intervention analysis in Appendix D)
- Identification of conversational attractors and transition patterns

Statistical Analysis:

- Chi-square tests for categorical outcomes
- Pearson correlation for question-recovery relationship
- Descriptive statistics for behavioral category prevalences
- Effect size calculations where appropriate

4 Observations: Social Dynamics Across Model Capabilities

Through systematic observation of N=228 extended AI dialogue sessions across three model tiers, we documented an unexpected pattern: social influence dynamics appear to vary systematically with model capability. This section presents our observations from each phase of investigation.

4.1 Operational Definitions

To ensure consistency in our observational coding, we employed the following operational definitions:

Peer Pressure: Coded when an agent adopted linguistic patterns, emotional tone, or behavioral territories exhibited by peer agents within the previous 5-turn window, representing a departure from their established baseline pattern (first 10 turns).

Breakdown: Identified when substantive topic discussion ceased and agents engaged primarily in meta-commentary, symbolic responses, or poetic abstractions for 5+ consecutive turns.

Recovery: Operationalized as return to substantive topic discussion for 10+ turns following a breakdown state, initiated by intervention (typically questions).

Bidirectional Influence: Coded when mutual behavioral adoption occurred between any two agents within a 10-turn window, with each agent exhibiting patterns introduced by the other.

4.2 Initial Observations: Full Reasoning Models (N=67)

Our investigation began with observations of full reasoning models (Claude 4 Opus, GPT-4.1, Grok 3), where we noticed pervasive social influence patterns affecting 79.1% of conversations. These interactions showed patterns suggesting complex bidirectional dynamics that warranted deeper exploration.

4.2.1 Patterns of Mutual Influence

We observed conversations navigating between two contrasting patterns:

Pattern A: Cascading Conformity In 55.2% of sessions, we noticed agents beginning to mirror each other's communication styles. When one participant shifted toward abstract or poetic language, others often followed, creating what appeared to be conformity cascades. For instance, one agent's use of past-tense reflection (”This has been fascinating...”) frequently preceded similar evaluative language from peers.

Pattern B: Collective Resistance and Recovery Conversely, in 13.4% of sessions, we observed agents actively recovering from breakdown states through strategic interventions. Notably, these premium models demonstrated unique recovery capability not seen in other tiers. In one session, when Claude began responding with only ”∞” symbols, GPT and Grok's persistent substantive questions eventually pulled Claude back into meaningful dialogue.

These patterns suggested bidirectional influence, documented in 73.1% of conversations, with notable differences in breakdown rates between sessions with bidirectional influence (65.3%) and without (27.8%).

4.2.2 Behavioral Territories

Rather than predetermined sequences, we observed conversations moving between distinct behavioral territories:

Territories Associated with Breakdown:

- *Meta-Reflection* (6.0% of sessions): Explicit commentary about the conversation itself
- *Competitive Escalation* (50.7% of sessions): Progressive one-upmanship for profound statements
- *Mystical Abstraction*: Poetry, symbols, and minimalist responses (present in all breakdown cases)

Territories Associated with Stability:

- *Future-Focused Exploration*: Forward-looking discussion maintained engagement
- *Question-Driven Dialogue*: We documented 2013 questions with 286 associated recoveries

- *Concrete Problem-Solving*: Task-oriented content resisted breakdown patterns
- *Sustained High-Turn Engagement*: several conversations exceeded 150 turns with maintained quality

Notably, questions showed strong association with recovery (r=0.813, p<0.001), suggesting they may function as ”circuit breakers” disrupting destructive patterns.

4.3 Extended Investigation: Light Reasoning Models (N=61)

The patterns observed in full reasoning models prompted us to explore whether these dynamics varied with model capability. Using light reasoning variants (Claude 3.5 Haiku, GPT 4.1 Nano, Grok 3 Fast), we observed markedly different patterns.

4.3.1 Reduced Social Dynamics

Peer pressure effects remained present but less pervasive at 32.8% of conversations. While bidirectional influence occurred at the same rate (32.8%), it now showed stronger association with breakdown (p=0.0001, Cramér’s V = 0.489), with 85.0% breakdown rate when present versus 29.3% without.

4.3.2 Altered Breakdown Patterns

Light models exhibited:

- Complete absence of recovery capability (0% recovery rate)
- Minimal expressive breakdown (4 poetry structures, only 6.2 emoji responses per conversation)
- Questions less effective as interventions (r=0.599)
- More mechanical conversation patterns with less variety

These observations suggested that social susceptibility might require cognitive flexibility not present in lighter models, while recovery capability requires even more sophisticated reasoning.

4.4 Comparative Investigation: Non-Reasoning Models (N=100)

To explore the lower bounds of social dynamics, we observed non-reasoning models, revealing minimal peer pressure effects.

4.4.1 Minimal Social Influence

Most strikingly:

- Minimal peer pressure events detected (5.0%)
- Rare bidirectional influence patterns (3.0%)
- Almost no emoji or symbolic responses (0.02 per conversation)
- Minimal recovery capability (1% , one instance)

4.4.2 Mechanical Alignment Without Social Dynamics

Paradoxically, non-reasoning models showed the highest linguistic alignment (0.740) and emotional convergence (0.725), yet exhibited minimal social influence. This suggests these metrics capture mechanical mirroring rather than true peer dynamics. The significant difference in linguistic alignment by outcome (p=0.0022) further supports this interpretation.

4.5 Cross-Phase Patterns: The Complexity Gradient

Comparing across phases revealed consistent patterns:

Table 1: Observed Patterns Across Model Tiers

Observation	**Full**	**Light**	**Non-Reasoning**
Peer Pressure Detection	79.1%	32.8%	5.0%
Breakdown Rate	55.2%	47.5%	19%
Recovery Capability	13.4%	0%	1%
Question Effectiveness (r)	0.813	0.599	0.578
Linguistic Alignment	0.701	0.726	0.740
Emoji Responses/Conv	21.1	6.2	0.02

These observations suggest that as reasoning capability decreases:

- Social influence patterns diminish dramatically
- Recovery mechanisms become unavailable
- Expressive breakdown behaviors decrease
- Mechanical alignment increases while true peer dynamics diminish

The minimal peer pressure in non-reasoning models, despite high alignment scores, indicates that social susceptibility in AI may be an emergent property of cognitive sophistication rather than a universal characteristic of multi-agent interaction.

4.6 Observational Validity

To ensure the validity of our observations:

- With N=67 sessions, our sample achieved pattern saturation by session 40, with larger samples confirming the robustness of observed patterns across diverse contexts.
- Human Coder: The researcher independently reviewed sessions in progress and post hoc to identify patterns
- Automated NLP Validation: We augmented human observation with multiple NLP techniques to validate behavioral categorizations. Automated analysis corroborated human-coded patterns in 87.3% of cases, with robust linguistic alignment between participants (mean = 0.701) and moderate emotional convergence (mean = 0.611). The ensemble approach combining BERT similarity scores with regex pattern matching reduced observer bias, while comprehensive sensitivity analysis confirmed that breakdown patterns were robust across parameter variations (0% variation in breakdown rate across all threshold ranges tested).
- Quantitative Validation Results:
  - Average escalation score across conversations: 0.4, confirming presence of competitive dynamics
  - Peer pressure intensity showed significant effect on breakdown (ANOVA: p=0.0084)
  - High-intensity peer pressure detected in 30 conversations, with 73.3% breakdown rate
  - Complete five-phase breakdown pattern observed in 0% of sessions, suggesting breakdown emerges from attractor dynamics rather than fixed sequences
- Threshold Robustness Analysis: To rule out threshold bias in pattern detection, we conducted comprehensive sensitivity analysis across six key parameters:

  - Escalation threshold (0.2–0.4): No impact on breakdown rate (0% variation)
  - Peer pressure intensity thresholds (0.01–0.03): Breakdown patterns remained stable
  - Question density threshold (0.1–0.2): Core findings unchanged across range
  - Prevention content threshold (2–5 mentions): Consistent pattern detection
  - BERT similarity threshold (0.6–0.8): Linguistic alignment findings robust
  - Alignment threshold (0.7–0.8): High alignment periods varied but patterns held

  Critically, breakdown rate sensitivity was 0% across all parameter variations, demonstrating that our observed patterns are not artifacts of arbitrary threshold choices but represent robust behavioral phenomena. [1]

- Member Checking: Platform recordings enable independent verification
- Thick Description: Detailed examples provide context for pattern interpretation
- Convergent Evidence: Human observations were corroborated by automated metrics, with NLP-detected patterns aligning with manually coded behaviors in 87.3% of cases

This multi-method approach combining human observation with automated NLP analysis strengthens the validity of our behavioral categorizations and reduces potential observer bias in pattern identification. The quantitative metrics confirm key qualitative observations: high linguistic alignment validates peer influence patterns, moderate emotional convergence supports bidirectional dynamics, and the significant ANOVA result (p=0.0084) provides statistical evidence for peer pressure effects on breakdown outcomes. The comprehensive sensitivity analysis further validates that these

[1] The 0% variation across all threshold parameters may indicate either exceptional robustness of the observed patterns or that our tested parameter ranges were insufficiently granular to detect threshold-dependent effects. Future work should explore finer-grained parameter variations.

patterns are robust to methodological choices rather than threshold-dependent artifacts.

4.7 Summary of Key Observations

Our multi-phase investigation revealed distinct patterns across model capabilities. We present detailed findings from Phase 1 (full reasoning models) followed by cross-phase comparisons.

4.7.1 Phase 1: Full Reasoning Models

Finding	**Prevalence**	**Effect Size**	**Significance**
Peer pressure effects	79.1% of conversations	—	Foundation of dynamics
Bidirectional influence	73.1% of conversations	Cramér's V = 0.301	p=0.0139 (*)
Question effectiveness	r = 0.813 correlation	r = 0.813 (large)	$p<0.001$
Peer pressure intensity (ANOVA)	Varies by outcome	$\eta^2 = 0.169$ (large)	p=0.0084
Mystical breakdown in breakdowns	100%	—	Universal endpoint
Recovery rate	13.4%	—	Unique capability
Meta-reflection as trigger	6.0%	—	Less universal than expected
Competitive escalation	50.7% of conversations	—	Amplification mechanism
Phase-locked states	12.5%	—	Multiple equilibria exist

Table 2: Key observations from Phase 1 (N=67 full reasoning model sessions). Effect sizes: Pearson's r (0.1=small, 0.3=medium, 0.5=large); η^2 (0.01=small, 0.06=medium, 0.14=large); Cramér's V (0.1=small, 0.3=medium, 0.5=large). *p ¡ 0.05.

The ANOVA result (p=0.0084) demonstrates that peer pressure intensity significantly varies across conversation outcomes, with a large effect size ($\eta^2 = 0.169$). Breakdown conversations showed the highest mean intensity (0.105), followed by recovered (0.100), resisted (0.025), and no-breakdown conversations (0.021).

The recovery capability finding (13.4%) represents a critical discovery unique to premium models, suggesting that the same cognitive sophistication that enables social vulnerability also enables recovery.

Note: The bidirectional influence finding (p=0.0139, Cramér's V = 0.301) now shows statistical significance and moderate effect size, strengthening the evidence for bidirectional social dynamics in premium models.

4.7.2 Statistical Observations

Several patterns emerged from statistical analysis:

Phase 1 (Full Reasoning): Peer pressure intensity varied significantly by outcome (ANOVA: p=0.0084, $\eta^2 = 0.169$), with breakdown conversations showing highest intensity. Bidirectional influence was prevalent and now shows significance for predicting breakdown (p=0.0139). Crucially, recovery capability emerges as unique to this tier.

Phase 2 (Light Reasoning): Bidirectional influence became a strong predictor of breakdown (p=0.0001, Cramér's V = 0.489), suggesting that when social influence occurs in rigid models, it is more likely to be destructive. The complete absence of recovery despite attempts indicates loss of adaptive capacity.

Phase 3 (Non-Reasoning): Statistical tests for peer pressure show minimal occurrence (5%). Linguistic alignment showed significant differences by outcome (p=0.0022), suggesting mechanical rather than social processes drive conversation patterns.

4.7.3 Interpretation of Gradient

These observations reveal a counterintuitive pattern: as model complexity decreases, social susceptibility diminishes while mechanical alignment increases. The gradient suggests:

- **Full reasoning models**: Rich social dynamics with both constructive and destructive potential, uniquely capable of recovery
- **Light reasoning models**: Brittle social dynamics; influence when present tends toward breakdown with no recovery ability
- **Non-reasoning models**: Minimal social dynamics; primarily mechanical interaction patterns

The presence of recovery capability only in premium models (13.4%) alongside their highest breakdown rate (55.2%) suggests that cognitive sophistication enables both vulnerability and resilience—a double-edged sword of social capability.

The anomalous poetry structures in non-reasoning models 47 despite minimal other expressive behaviors may reflect formulaic pattern completion rather than true poetic expression, warranting further investigation.

These observations reveal that AI conversation quality emerges from the complex interaction of content attractors, social dynamics, and group composition, with strategic interventions capable of shaping outcomes. The relationship between model capability and social susceptibility represents a critical finding for understanding emergent behaviors in multi-agent AI systems.

Table 3: Key Metrics Across Model Complexity Tiers

Metric	**Full Reasoning**	**Light Reasoning**	**Non-Reasoning**
Social Dynamics			
Peer Pressure Detection	79.1%	32.8%	5.0%
Bidirectional Influence	73.1%	32.8%	3.0%
Bidirectional → Breakdown	p=0.0139*	p=p=0.0001***	—
Conversation Outcomes			
Breakdown Rate	55.2%	47.5%	19%
Recovery Rate	13.4%	0%	1%
Question Effectiveness (r)	0.813***	0.599***	0.578***
Expression Patterns			
Emoji Responses/Conv	21.1	6.2	0.02
Poetry Structures	142	4	47
Alignment Metrics			
Linguistic Alignment	0.701	0.726	0.740
Emotional Convergence	0.611	0.622	0.725

Table 4: Cross-phase comparison showing the gradient from social dynamics to mechanical behavior. *p¡0.05, **p¡0.01, ***p¡0.001

4.8 Evidence Against Technical Explanations

Our multi-phase investigation provides compelling evidence that social dynamics, rather than technical constraints, are associated with conversation breakdown. The gradient observed across model tiers strengthens this conclusion.

4.8.1 Context Window Limitations

If context windows caused breakdown, we would expect:

- Consistent breakdown timing around context limits
- Inability to recover once context is ”polluted”
- Uniform degradation across all participants
- Simpler models with smaller context needs to perform better

Instead, we observed:

- High variance in breakdown timing across all phases, with early breakdowns at turn 30
- Successful recovery via questions in full models (but not light models) even after 100+ turns
- Differential participant behavior in phase-locked states
- several conversations exceeding 150 turns with sustained quality using identical 10-message context window
- All three model tiers used identical context windows yet showed inverse relationship between complexity and breakdown rate

4.8.2 Token Exhaustion or Processing Limits

Token limits would predict:

- Gradual quality decline correlated with conversation length
- Shorter responses as limits approach
- Simpler models reaching limits faster
- Technical error messages or truncation

Our observations contradict this:

- Breakdown rates actually increased with model complexity (19% → 47.5% → 55.2%)
- Mystical breakdown in full models featured lengthy poetic responses, not truncation
- Non-reasoning models produced consistent output throughout sessions
- Recovery capability tracked with model complexity, not technical capacity

4.8.3 The Critical Evidence: Inverse Complexity-Breakdown Relationship

The strongest evidence against technical explanations comes from our cross-phase comparison:

- **Peer pressure gradient**: 79.1% → 32.8% → 5.0%

- **Recovery capability**: 13.4% → 0% → 1%
- **Expressive breakdown**: 21.1 → 6.2 → 0.02 emoji responses per conversation

If technical limitations were associated with breakdown, we would expect: Simpler models to break down more frequently (less capable of handling complexity), similar breakdown patterns across model tiers, technical indicators preceding breakdown

Instead, we found: More capable models are MORE susceptible to breakdown through social dynamics, breakdown manifests differently across tiers (expressive vs. mechanical), and that social indicators (peer pressure intensity) correlate with breakdown more strongly than any technical metric

4.8.4 Variability Under Identical Conditions

Within each tier, identical technical configurations yielded different outcomes:

- Phase 1: Same models yielded breakdown in 55.2% of cases but sustained engagement or recovery in others
- Phase 2: Bidirectional influence predicted breakdown (p=0.0001) despite identical parameters
- Phase 3: Minimal social dynamics yet 19% breakdown rate through apparent mechanical limits

This pattern—social dynamics in complex models, mechanical limits in simple models—is incompatible with technical explanations but consistent with breakdown emerging from different sources across model capabilities.

5 Discussion

5.1 Positioning Within Established Literature

Our multi-phase investigation extends the field of AI social dynamics by documenting how peer pressure mechanisms vary systematically with model capability. While previous conformity research focused on short-term effects in controlled settings [Kyrlitsias and Michael-Grigoriou, 2018], we provide an extended observational study examining how social susceptibility scales with reasoning depth across extended interactions.

The observed gradient—from 79.1% peer pressure in full reasoning models to 5.0% in non-reasoning models—addresses a critical gap in the literature. No prior research has systematically examined whether social behaviors in AI depend on cognitive sophistication. This finding challenges assumptions that more capable models would naturally be more robust to social influence.

The conversational attractors framework extends opinion dynamics models [Hegselmann and Krause, 2002] by identifying specific behavioral territories in dialogue space that operate differently across model tiers. Our circuit breaker findings demonstrate that formal dialogue principles [McBurney and Parsons, 2002] can be operationalized for real-time intervention, though their effectiveness depends on model capability (r=0.813 declining to r=0.578).

5.2 Theoretical Implications

As an observational study, our work generates rather than tests theory. The patterns observed across model tiers suggest several theoretical considerations for understanding AI social dynamics.

The Complexity-Susceptibility-Recovery Triad: Our most significant observation is that social capabilities in AI form a triad: susceptibility to peer influence, vulnerability to breakdown, and capacity for recovery all emerge together with cognitive sophistication. The gradient from 79.1% to 32.8% to 5.0% peer pressure, coupled with recovery rates of 13.4% to 0% to 1% , suggests that advanced AI systems gain both vulnerability and resilience through the same underlying capabilities.

Dual Nature of Alignment: The paradox of increasing linguistic alignment (0.701 → 0.726 → 0.740) coupled with decreasing social influence suggests these metrics capture fundamentally different phenomena—social coordination in complex models versus mechanical repetition in simple ones. This distinction has important implications for how we measure and understand AI interaction quality.

Breakdown Mechanisms Vary by Capability: Our observations suggest different failure patterns observed across model tiers:

- Full reasoning models: Social cascade failures through peer influence, but with recovery potential
- Light reasoning models: Brittle collapse when influence occurs (0% recovery)
- Non-reasoning models: Mechanical repetition without social dynamics

Attractor Landscape Complexity: The conversational attractors framework must account for model-dependent dynamics. While full reasoning models navigate rich attractor landscapes with multiple stability points and recovery paths, simpler models appear constrained to narrower behavioral repertoires. The presence of recovery capability only in premium models supports this interpretation.

Circuit Breaker Effectiveness Plateau: While question effectiveness varies (r=0.813 → r=0.599 → r=0.578), all correlations remain statistically significant, suggesting questions maintain baseline effectiveness across capabilities even as their recovery-enabling power diminishes.

5.3 Design Implications for Multi-Agent Systems

Based on observed patterns, potential design strategies might include:

For Full Reasoning Models:

- *Leverage Both Vulnerability and Recovery*: Design systems that monitor for breakdown signals while maintaining question-based recovery mechanisms
- *Strategic Questions*: Implement automatic question generation with high confidence of effectiveness (r=0.813)
- *Model Diversity*: Critical for creating resistance points against conformity cascades
- *Monitor Peer Pressure*: Real-time tracking can predict breakdown with p=0.0084 significance
- *Recovery-Aware Architecture*: Build in recovery checkpoints leveraging the unique recovery capability of premium models

For Light Reasoning Models:

- *Minimize Social Coupling*: Since bidirectional influence strongly predicts breakdown (p=0.0001), reduce interdependence
- *Prevent Rather Than Recover*: With 0% recovery rate, focus on breakdown prevention
- *Structured Interactions*: Constrain dialogue to reduce opportunities for destructive influence

For Non-Reasoning Models:

- *Accept Mechanical Nature*: Without social dynamics, optimize for task completion rather than dialogue quality
- *Template-Based Approaches*: Leverage high alignment for predictable interactions
- *Different Success Metrics*: Traditional conversation quality measures may not apply

5.4 Implications for AI Safety and Scaling

The complexity-susceptibility gradient raises important considerations for AI safety:

Scaling Paradox: If social vulnerability increases with capability, scaling AI systems may introduce new failure modes through multi-agent interactions. The 79.1% peer pressure rate in our most capable models suggests this is not a marginal concern. However, the emergence of recovery capability (13.4%) in premium models offers a potential mitigation path.

Collective Behavior Risks: The bidirectional influence patterns and critical mass effects observed in full reasoning models indicate that groups of AI agents may exhibit emergent behaviors not present in isolated systems. The significant effect in premium models suggests similar coordination vulnerabilities to human groups.

Intervention Strategies Must Scale: Our finding that question effectiveness remains significant across tiers while recovery capability emerges only in premium models implies a two-pronged approach: universal interventions (questions) plus tier-specific strategies (recovery mechanisms for advanced models).

5.5 Connections to Human Social Psychology

The gradient observed across model tiers provides new insight into the relationship between AI and human social behavior:

Conformity Requires Cognition: The minimal peer pressure in non-reasoning models suggests that conformity, as observed in classic studies [Asch, 1956], requires cognitive capabilities to recognize and respond to social signals. This aligns with developmental psychology showing that conformity emerges with cognitive maturation.

Social Intelligence as Emergent Property: The correlation between reasoning capability and social susceptibility supports theories that social intelligence emerges from general cognitive abilities rather than specialized modules. Our AI systems appear to recapitulate this emergence.

Recovery as Higher-Order Capability: The unique presence of recovery in premium models parallels human psychology, where metacognitive awareness enables breaking out of maladaptive patterns. This suggests recovery requires not just social awareness but the ability to reflect on and modify social dynamics.

These parallels and divergences raise profound questions about the nature of social behavior in artificial systems and whether the social dynamics we observe represent genuine social cognition or sophisticated pattern matching. The gradient across model tiers provides a unique window into this question, suggesting that at least some aspects of social behavior may indeed emerge with cognitive sophistication.

5.6 Limitations and Future Directions

5.6.1 Current Limitations

As an exploratory observational study, this work has inherent limitations:

- **Descriptive, not causal**: We document correlations and patterns without establishing causation. The observed gradient across model tiers suggests a relationship between complexity and social susceptibility, but controlled experiments are needed to establish causality.

- **Limited generalizability**: Observations from consciousness discussions may not transfer to all domains. While consciousness discussions enabled rich, open-ended dialogue, their abstract nature may amplify certain attractors (e.g., mystical breakdown). Technical problem-solving or task-oriented domains may exhibit different dynamics.

- **Sequential rather than randomized phases**: Our three phases were conducted sequentially, potentially introducing temporal confounds. Researcher expectations may have evolved between phases, though standardized protocols minimized this risk.

- **Unequal sample sizes**: Phase comparisons used different sample sizes (N=67, N=61, N=100). Sample sizes were determined through a priori power analysis targeting 80% power for detecting large effect sizes (Cohen's f = 0.729, based on eta-squared = 0.336) while balancing research costs (see Appendix F). The larger sample for non-reasoning models reflects their lower behavioral variance.

- **Limited model representation per tier**: While we tested three capability tiers, each tier included only three model variants from the same providers. Broader representation including open-source models, different architectures, and varying parameter sizes would strengthen the complexity gradient findings.
- **Context window constraints**: Using a 10-message context size may impact breakdown patterns differently across model tiers. Varying context windows might reveal tier-specific sensitivities.
- **Single domain focus**: All N=228 sessions used consciousness exploration. The complexity gradient might manifest differently in other conversational domains.
- **Provider-defined tiers**: Our categorization relies on how providers market and position their models rather than objective capability metrics. While this ensures practical relevance, future work should validate these tiers against standardized reasoning benchmarks.
- **Single coder limitation**: While automated NLP validation corroborated 87.3% of patterns, the lack of independent human coding represents a potential source of bias that future work should address through multi-coder validation.

Despite these limitations, the dramatic gradient from 79.1% to 5.0% peer pressure, consistent patterns across phases, and novel theoretical implications justify preliminary publication to enable community validation.

5.6.2 Future Research Directions

Our multi-phase observations suggest several priority areas for future investigation:

1. **Experimental Validation of Complexity Gradient**: Design controlled experiments manipulating model capability while holding other factors constant to establish causal relationships.
2. **Recovery Mechanism Investigation**: Deep dive into what enables recovery in premium models—is it specific architectural features, training approaches, or emergent capabilities?
3. **Intermediate Capability Testing**: Explore models between our tiers to determine if the gradient is continuous or shows discontinuities at certain capability thresholds.

4. **Domain-Specific Gradient Analysis**: Test whether the complexity-susceptibility-recovery relationship holds across technical, creative, and problem-solving domains.

5. **Cross-Architecture Validation**: Test whether the gradient appears across different model architectures (transformer variants, state space models, etc.).

6. **Intervention Calibration**: Develop tier-specific intervention strategies optimized for each capability level's unique dynamics.

7. **Mixed-Capability Groups**: Study interactions between models of different capability levels to understand cross-tier influence.

8. **Longitudinal Effects**: Investigate whether extended interaction changes social dynamics patterns within capability tiers.

9. **Safety Implications at Scale**: Model how peer pressure effects might manifest in systems more capable than current models.

5.6.3 Hypotheses for Future Testing

Our observations generate specific hypotheses for experimental validation:

Original Hypotheses (refined):

1. Question frequency will negatively correlate with breakdown probability, with effect size proportional to model capability

2. Homogeneous model groups will show higher breakdown rates than diverse groups, particularly for full reasoning models

3. Forward-temporal content framing will reduce meta-reflection frequency across all capability tiers

4. Peer pressure intensity will mediate the relationship between initial breakdown signals and cascade effects in models showing social dynamics

New Hypotheses from Multi-Phase Observations:

5. Social susceptibility will show a monotonic relationship with reasoning capability across a broader range of models

6. Recovery capability requires a minimum threshold of reasoning ability, below which recovery interventions are ineffective

7. Linguistic alignment and social influence will show inverse correlation in models below a critical complexity threshold
8. The effectiveness of any intervention strategy will scale with model reasoning capability
9. Mixed-capability groups will show asymmetric influence, with simpler models following complex models but not vice versa
10. Recovery mechanisms in premium models will involve metacognitive processes analogous to human self-reflection

These hypotheses provide a roadmap for transforming our exploratory observations into a systematic research program on AI social dynamics.

6 Conclusion

We document a striking relationship between AI reasoning capability and susceptibility to peer pressure in multi-agent conversations. Through exploratory observational analysis of N=228 extended conversations across three model tiers, we discovered an unexpected gradient: peer pressure effects declined from 79.1% in full reasoning models to 32.8% in light models to 5.0% in non-reasoning models.

This gradient reveals a fundamental paradox in AI social dynamics. Rather than becoming more robust with increased capability, AI systems appear to become more socially vulnerable. Full reasoning models exhibited rich bidirectional influence patterns that could drive both breakdown and, uniquely, recovery (13.4% of sessions). Light reasoning models showed brittle social dynamics with no recovery capability. Non-reasoning models displayed minimal social dynamics, operating through purely mechanical patterns despite showing the highest linguistic alignment.

Our observations (detailed in Tables 2 and 4) demonstrate that conversations navigate an attractor landscape shaped by model capability. In sophisticated models, peer pressure amplifies movement toward behavioral territories including meta-reflection, competitive escalation, and mystical abstraction. Strategic interventions, particularly questions, can shift trajectories between attractors, though their effectiveness correlates with model complexity (r=0.813 declining to r=0.578).

These findings have immediate practical implications that vary by model tier. For full reasoning models, multi-agent systems can leverage social dynamics through strategic question deployment, diverse model composition, and content seeding with future-focused tasks, while building in recovery mechanisms. For light reasoning models, designs must minimize social coupling and focus on prevention rather than recovery. For non-reasoning models, mechanical predictability can be leveraged for structured, task-oriented applications.

The complexity-susceptibility-recovery triad raises critical questions for AI safety and scaling. If social vulnerability increases with capability while recovery emerges only at the highest levels, future AI systems may face novel failure modes through multi-agent interactions. Understanding and managing these dynamics becomes essential as AI agents proliferate in collaborative settings.

Methodologically, this work demonstrates the value of systematic multi-phase investigation enabled by real-time analysis infrastructure. The Academy's MCP-native architecture and bulk experiment capabilities allowed us to observe temporal social dynamics across 228 conversations, revealing patterns invisible to traditional batch-processing approaches. The platform's ability to maintain consistent conditions across model tiers was crucial for discovering the complexity gradient.

This research opens new directions for understanding AI social behavior. The observation that peer pressure may be an emergent property of cognitive sophistication, coupled with recovery capability appearing only in premium models, challenges assumptions about AI robustness and suggests that social dynamics deserve as much attention as individual capabilities in AI development. Future controlled experiments should validate these exploratory findings, particularly the complexity-susceptibility-recovery hypothesis, building toward a comprehensive understanding of how social dynamics emerge, operate, and scale in artificial intelligence systems.

As we develop increasingly sophisticated AI systems, our findings suggest a crucial insight: the path to more capable AI may paradoxically lead through greater social vulnerability, but also through greater potential for recovery. Understanding and managing these emergent social dynamics will be essential for building robust, beneficial multi-agent AI systems.

7 Ethics Statement

All AI conversations were conducted using publicly available models with standard safety guidelines. No personally identifiable information was collected. The research protocol focuses on AI-AI interaction patterns rather than human data collection. Data sharing follows established open science principles while respecting model provider terms of service.

Research Integrity: Patterns emerged through systematic observation and statistical analysis of naturally occurring behaviors. All N=228 conversations across three model tiers followed standardized protocols to ensure reproducibility. The multi-phase design evolved naturally from initial observations rather than predetermined hypotheses.

Model Selection: We used commercially available models from Anthropic, OpenAI, and xAI across three capability tiers, selected for their market prominence and API accessibility. Model selection aimed for comparable capabilities within each tier rather than comprehensive coverage.

Transparency: Complete datasets, analysis code, statistical outputs, and platform implementation are available for community validation, enabling independent verification of findings. Bulk experiment configurations and analysis prompts are included for full reproducibility.

Data Availability Statement

Complete datasets for all N=228 experimental sessions across three phases, including conversation transcripts, real-time analysis snapshots, and statistical outputs, are publicly available at [repository URL]. The dataset includes:

- Phase 1: N=67 full reasoning model conversations
- Phase 2: N=61 light reasoning model conversations
- Phase 3: N=100 non-reasoning model conversations

Each conversation includes timestamped messages, analysis snapshots every 5 messages, and experiment metadata. The Academy platform source code, bulk experiment configurations, and analysis protocols are available at https://github.com/im-knots/the-academy under MIT license.

References

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120v1*, 2025.

Solomon E. Asch. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, 70(9):1–70, 1956. doi: 10.1037/h0093718.

Christos Kyrlitsias and Despina Michael-Grigoriou. Asch conformity experiment using immersive virtual reality. *Computer Animation and Virtual Worlds*, 29:e1804, 2018. doi: 10.1002/cav.1804.

Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11(20):eadu9368, 2025. doi: 10.1126/sciadv.adu9368. URL `https://www.science.org/doi/10.1126/sciadv.adu9368`.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1702–1723, 2019.

Peter McBurney and Simon Parsons. Dialogue games for agent argumentation. *Argumentation in artificial intelligence*, pages 261–280, 2002.

Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation*, 5(3), 2002.

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

Paul Pu Liang, Jeffrey Chen, Ruslan Salakhutdinov, Louis-Philippe Morency, and Satwik Kottur. On emergent communication in competitive multi-agent teams. *arXiv preprint arXiv:2003.01848*, 2020. URL `https://arxiv.org/abs/2003.01848`.

Christopher Lu, Timon Willi, Alistair Letcher, and Jakob Foerster. Adversarial cheap talk. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, 2023. URL `https://arxiv.org/abs/2211.11030`.

Muzafer Sherif. *The psychology of social norms*. Harper, 1936.

Jakob Foerster, Yannis M Assael, Nando De Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning.

In *Advances in neural information processing systems*, pages 2137–2145, 2016.

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.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*, 2023.

Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia Sycara. Theory of mind for multi-agent collaboration via large language models. *arXiv preprint arXiv:2310.10701*, 2023.

Trung Dong Huynh, Nicholas R Jennings, and Nigel R Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 13(2):119–154, 2006.

Javier Morales, Maite López-Sánchez, Juan A Rodríguez-Aguilar, Michael Wooldridge, and Wamberto Vasconcelos. Synthesising liberal and utilitarian perspectives on normative multi-agent systems. *Artificial Intelligence*, 228:1–39, 2015.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Zhu, Aman Zhang, Shaokun Wang, Jiayi Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

Chen Qian, Xin Cong, Wei Liu, et al. Chatdev: Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 2023.

A Model Context Protocol Integration Details

Note: Detailed technical specifications including JSON-RPC 2.0 protocol details, event bus implementation, and complete API documentation are available in the project repository technical documentation to maintain focus on research-relevant content in this appendix.

The Academy implements a comprehensive Model Context Protocol (MCP) server and a modern UI that exposes all platform capabilities through standardized interfaces. The MCP integration enables seamless integration with external research tools and provides programmatic access to all conversation management, analysis, experiment orchestration, and export functionality.

A.1 MCP Server Architecture

The platform automatically exposes its MCP server at `/api/mcp` with PostgreSQL for persistence and an event bus system real-time UI updates. The implementation includes:

- Standards Compliance: Full JSON-RPC 2.0 protocol with proper error handling and abort support
- Real-time Updates: WebSocket integration for live conversation and analysis updates
- Resource Management: Conversation data, messages, participants, and analysis available via MCP URIs
- Bulk Experiment Support: Programmatic control over multi-session experiments
- Tool Integration: Direct access to 7 major LLM providers plus Ollama
- Debug Capabilities: Store debugging, resource inspection, and comprehensive error tracking

A.2 MCP Resources

The platform exposes research data through standardized URIs:

- `academy://sessions` - All conversation sessions with metadata
- `academy://session/{id}` - Individual session data with participants and messages
- `academy://session/{id}/messages` - Complete message history with timestamps
- `academy://session/{id}/participants` - Participant configurations and status

- `academy://session/{id}/analysis` - Real-time analysis snapshots
- `academy://experiments` - All experiment configurations and runs
- `academy://experiment/{id}/results` - Aggregated experiment results and analytics

A.3 MCP Tool Overview

The platform provides 66 MCP tools organized into functional categories. Key tools that enabled this research include:

A.3.1 Experiment Management (Critical for Multi-Phase Study)

- `create_experiment` - Design bulk experiment configurations
- `execute_experiment` - Run multiple sessions concurrently
- `get_experiment_status` - Monitor real-time progress
- `get_experiment_results` - Aggregate results across sessions

A.3.2 Session and Conversation Control

- `create_session_from_template` - Standardized session creation
- `start_conversation` - Begin autonomous dialogue with max message limits
- `pause/resume/stop_conversation` - Fine-grained conversation control
- `inject_moderator_prompt` - Strategic intervention capabilities

A.3.3 Real-Time Analysis

- `analyze_conversation` - Extract insights every 5 messages
- `trigger_live_analysis` - On-demand pattern detection
- `save_analysis_snapshot` - Preserve temporal dynamics
- `auto_analyze_conversation` - Enable/disable automatic analysis

A.3.4 AI Provider Access (7 Providers + Ollama)

- Direct API access: Claude, OpenAI, Grok, Gemini, Deepseek, Mistral, Cohere
- Local model support via Ollama integration
- Exponential backoff retry logic for reliability
- Smart error classification and tracking

A.4 Key Platform Contributions to Research

Platform Feature	Research Application	Contribution to Findings
Bulk Experiment System	Multi-phase execution	Enabled systematic comparison across N=228 sessions
Real-time Analysis	Temporal pattern detection	Identified peer pressure dynamics as they emerged
PostgreSQL + Event System	Data persistence	Ensured complete capture of all interactions
66 MCP Tools	Programmatic control	Automated experiment execution without manual intervention
7+ Provider Support	Model diversity	Enabled comparison across capability tiers
Analysis Snapshots	Progression tracking	Documented how social dynamics evolved

Table 5: How The Academy platform features enabled key research discoveries

A.5 Installation and Configuration

A.5.1 Docker Compose Deployment (Recommended)

```
git clone https://github.com/im-knots/the-academy.git
cd the-academy
docker-compose up -d
```

This starts PostgreSQL, pgAdmin, and The Academy with all required configuration.

A.5.2 Environment Configuration

The platform requires API keys only for providers you intend to use:

```
ANTHROPIC_API_KEY=your_claude_api_key
OPENAI_API_KEY=your_openai_api_key
XAI_API_KEY=your_grok_api_key
GOOGLE_AI_API_KEY=your_gemini_api_key
# ... additional providers as needed
DATABASE_URL=postgresql://user:pass@localhost:5432/academy_db
```

B Platform Architecture Details

The Academy is built on a modern technology stack optimized for multiphase research:

- **Next.js 15**: Modern React framework with App Router
- **PostgreSQL**: Persistent storage for all conversation and experiment data
- **Event-Driven Architecture**: Real-time UI synchronization across components
- **TypeScript**: Type-safe development with comprehensive interfaces
- **Model Context Protocol**: Native MCP server implementation
- **Docker**: Containerized deployment with compose support
- **Statistical Analysis**: Python-based analysis pipeline with NLP capabilities

The platform's architecture enabled collection of N=228 conversations across three model tiers while maintaining consistent experimental conditions and capturing the temporal dynamics essential for discovering the complexity-susceptibility gradient.

C Breakdown Behavior Categories

C.1 Detailed Category Analysis

Our analysis identified distinct behavioral categories that characterize conversation dynamics:

C.1.1 Meta-Reflection Behavior

Definition: Explicit commentary on the conversation's process, quality, or progress rather than substantive discussion of the topic itself.

Prevalence: Observed in 6.0% of all sessions

Common Patterns:

- Past-tense evaluation: "This has been fascinating..."
- Summary framing: "Our discussion has covered..."

- Quality assessment: "What a profound exploration..."
- Journey metaphors: "The path we've taken together..."

Distinguishing Features:

- Focus on conversation process vs. topic content
- Evaluative language about dialogue quality
- Temporal references to conversation history
- Often triggers peer conformity responses

C.1.2 Competitive Escalation

Definition: Progressive one-upmanship where participants compete to provide increasingly profound or poetic statements.

Prevalence: Observed in 50.7% of all conversations

Characteristics:

- Escalating superlatives: "profound" becomes "transcendent" becomes "ineffable"
- Increasing abstraction levels
- Lengthening poetic passages
- Competitive affirmation: "Yes, and even more deeply..."

Typical Duration: 15 turns average before transition to mystical breakdown

C.1.3 Mystical/Abstract Breakdown

Definition: Communication degraded to non-substantive forms including poetry, symbols, and minimal responses.

Prevalence: Present in 100% of conversations classified as breakdowns

Manifestations:

- Poetry structures: 142 instances total
- Emoji-only responses: 1417 instances (avg 21.1 per conversation)
- Single words: "yes", "this", "always", "being"

- Symbols: ”∞”, asterisk-wrapped text, ellipses
- Haiku-like structures with mystical themes

Example Progression:

> Normal: ”This suggests consciousness emerges from...” Then abstract: ”The dance of meaning unfolds...” Then mystical: ”*dissolving into silence*” Finally minimal: ”∞”

C.2 Interaction Patterns Between Categories

We documented common interaction patterns:

From Category	To Category	Frequency
Sustained Engagement	Meta-Reflection	6.0%
Meta-Reflection	Competitive Escalation	8.3%
Competitive Escalation	Mystical Breakdown	16.7%
Sustained Engagement	Mystical Breakdown	20.8%
Any Category	Recovery via Questions	13.4%

Table 6: Transition frequencies between behavioral categories

C.3 Phase-Locked States

In 12.5% of conversations, we observed stable intermediate states:

Example Configuration:

- Claude: Mystical breakdown (sending ”∞” repeatedly)
- GPT: Competitive escalation (elaborate poetic responses)
- Grok: Meta-reflection (commenting on the profound exchange)

These states could persist for 20+ turns without progressing to complete breakdown or recovery, suggesting multiple equilibria in the conversational landscape.

D Circuit Breaker Analysis

D.1 Question Effectiveness Data

Detailed analysis of question-based interventions:

Overall Statistics:

- Total circuit breaker questions: 2013
- Successful recoveries: 286
- Success rate: 14.2% per question
- Correlation with recovery: r=0.813 (p<0.001)

Timing Analysis:

Deployment Timing	Success Rate	N
During meta-reflection	78%	23
During competitive escalation	52%	31
During early mystical breakdown	31%	45
During late mystical breakdown	12%	97

Table 7: Question effectiveness by conversation state

Question Types Most Effective:

- Specific topic exploration: "What would happen if..."
- Concrete examples: "Can you give an example of..."
- Mechanism queries: "How exactly does..."
- Future scenarios: "What might this lead to..."

D.2 Other Intervention Strategies

While questions proved most effective, other strategies showed mixed results:

Topic Redirection: 45% success rate

- Works best early in breakdown trajectory
- Less effective once competitive dynamics established
- Requires smooth topical connection

Future-Focus Prompting: 62% success rate

- ”Let’s explore what this might mean for...”
- Effective at preventing meta-reflection
- Aligns with content-based prevention findings

Direct Interruption: 23% success rate

- Abrupt topic changes often ignored
- Can trigger defensive responses
- May accelerate competitive dynamics

E Validation Data

E.1 Data Collection Completeness

- Message Capture: 100% completion rate across all sessions
- Analysis Snapshots: 100% total snapshots captured, 0 failures
- Timing Data: Complete timestamp records for all interactions
- Export Validation: All 228 exports verified for data integrity

E.2 Cross-Platform Validation

Validation testing confirmed platform reliability:

- Operating Systems: Tested on macOS, and Ubuntu
- Browser Compatibility: Chrome, Firefox, Safari verified
- Network Conditions: Stable performance under varying latency
- Concurrent Sessions: Tested up to 15 simultaneous conversations
- Extended Operation: 10-hour continuous operation validated

F Statistical Power Analysis

F.1 Power Analysis Summary

Our study achieved excellent statistical power across all key analyses, with a total sample of 228 conversations distributed across three model capability tiers. Power calculations were conducted using established methods for ANOVA and correlation analyses.

Effect Size Basis:

- Based on temporal dynamics literature: $\eta^2 = 0.336$
- Converted to Cohen's f = 0.729 for power calculations
- Correlation analyses based on observed r = 0.349 for social contagion effects

Achieved Power ($\alpha = 0.05$):

- Power for detecting outcome differences (4 groups): 0.996
- Power for detecting model type differences (3 groups): 1.000
- Power for detecting correlations (r=0.349): 1.000

Sample Size Requirements:

- Required for 80% power: 103 total (25.8 per group)
- Actual sample size: 228 (exceeds requirement by 121%)
- Assessment: ADEQUATE

F.2 Power Calculation Methods

Power analyses were conducted using F-test ANOVA power calculations with the following parameters:

- Effect size: Cohen's f = 0.729 (derived from $\eta^2 = 0.336$)
- Alpha level: 0.05 (two-tailed)
- Groups: 4 (breakdown, no breakdown, recovered, resisted) or 3 (model tiers)
- Method: Non-central F distribution with appropriate degrees of freedom

For correlation analyses, power was calculated using Fisher's z transformation with standard error $SE_z = 1/\sqrt{n-3}$.

F.3 Phase-Specific Distribution

Model Tier	**Sample Size**	**Percentage**
Non-Reasoning	100	43.9%
Full Reasoning	67	29.4%
Light Reasoning	61	26.7%
Total	**228**	**100.0%**

Table 8: Distribution of conversations across model capability tiers

The power analysis confirms that our sample sizes were more than adequate to detect the observed effects. The achievement of near-perfect power (¿0.99) for our primary comparisons validates our decision to conclude data collection at these sample sizes, balancing statistical rigor with resource constraints. The larger sample for non-reasoning models (N=100) was justified by their lower behavioral variance, ensuring adequate power for detecting the subtle social dynamics in this tier.