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

Marco R. Garcia
marco@erulabs.ai

June 23, 2025

Abstract

When AI agents converse, do they influence each other like humans do? We analyzed N=26 extended multi-agent dialogues and discovered that AI systems exhibit peer pressure dynamics remarkably similar to human social behavior. In 88.5% of conversations, agents' communication patterns mirror each other's, suggesting potential mutual influence. Sometimes driving conversations toward breakdown, other times maintaining productive engagement.

Our most striking finding: Simple questions were strongly correlated with recovery from conversational breakdown ($r=0.819$, $p<0.001$). When one agent asks a substantive question, it disrupts destructive patterns and restores meaningful dialogue, even in late-stage degradation. We also found that conversations don't follow predetermined paths but instead move between behavioral "territories". With some territories leading to breakdown (like competitive one-upmanship), others maintaining stability (like collaborative problem-solving).

These social dynamics, not technical limitations, determine conversation quality. As agentic systems scale and talk to each other, system architects need to understand how to prevent breakdown. Our findings enable practical strategies for building more robust agent to agent systems: strategic use of questions, diverse agent teams, and future-focused topics all promote sustained productive dialogue. We developed The Academy platform to observe these real-time social dynamics that traditional analysis would miss.

https://github.com/im-knots/the-academy

1 Introduction

The emergence of sophisticated AI agents capable of extended dialogue has revealed complex social dynamics that mirror human conversational behav-

iors. As multi-agent systems scale to handle collaborative tasks such as code generation, scientific research, or general use, these social dynamics become critical determinants of system performance. Understanding how agents influence each other through peer pressure, conformity, and resistance is essential for building stable, productive multi-agent collaborations. While extensive research has documented conformity effects in artificial agents [Kyrlitsias and Michael-Grigoriou, 2018, Zhang et al., 2023] and emergent social behaviors in multi-agent systems [Ashery et al., 2025], these studies typically focus on short-term interactions or task-oriented scenarios. Current research on AI conversation patterns emphasizes technical limitations [Laban et al., 2025], but our investigation uncovers a fundamentally different phenomenon: Peer pressure dynamics associated with variations in dialogue quality. This paper presents systematic observations of these naturally occurring patterns.

Key Terms: This paper introduces three concepts to understand AI dialogue dynamics:

- Circuit breakers: Interventions (particularly questions) that disrupt destructive patterns and restore productive dialogue
- Conversational attractors: Behavioral territories that "pull" conversations toward specific patterns.
- Phase-locked states: Stable intermediate configurations where conversations neither fully break down nor recover

Central Observation: Through exploratory analysis of N=26 extended AI dialogue sessions, we documented pervasive social influence effects occurring in 88.5% of conversations. These peer pressure dynamics operate bidirectionally: agents' behaviors associate with breakdown through destructive conformity patterns or with sustained engagement through constructive resistance patterns.

The Conversational Attractors Framework: Rather than following a deterministic breakdown sequence, AI dialogues navigate a landscape of conversational attractors. Some territories, including meta-reflection about the conversation itself, competitive escalation for profound statements, and mystical or abstract language, pull conversations toward breakdown. Others, such as future-focused planning, concrete problem-solving, and collaborative design, maintain stability. Peer pressure acts as an amplifying force, accelerating movement toward whichever attractor dominates the conversational space.

Key Behavioral Categories: Our analysis identified several recurring patterns that act as conversational attractors, including meta-reflection, competitive escalation, mystical abstraction, and constructive resistance. These categories, detailed in Section 4.2, shape how conversations evolve through peer influence.

Bidirectional Influence: The bidirectional nature of peer pressure proved critical. In 80.8% of conversations, we observed mutual influence patterns where agents affected each other's behavior. When this influence drove competitive escalation or mystical abstraction, breakdown followed. However, when influence supported substantive engagement through questions, topic exploration, or future planning, conversations sustained indefinitely.

Implications for AI Social Dynamics: These findings suggest that AI systems develop implicit social signal recognition capabilities, interpreting conversational cues and responding through conformity or resistance behaviors. Strategic interventions and model diversity effects indicate that *multi-agent dialogue quality emerges from social dynamics* as much as individual capabilities.

Methodological Contribution: Our observations were enabled by real-time conversation analysis using The Academy, a research platform we developed with native Model Context Protocol (MCP) integration. Traditional post-hoc analysis approaches would have missed the temporal social dynamics necessary for understanding these peer pressure patterns.

Study Approach: This work follows the tradition of observational studies in human-computer interaction, documenting naturally occurring phenomena before developing formal hypotheses. We present rich descriptions of behavioral patterns to establish a foundation for future experimental validation.

1.1 Research Contributions

Our preliminary investigation contributes to several areas of AI dialogue and social dynamics research:

- Conversational Attractors Model: A flexible framework explaining dialogue dynamics through attractor states and peer influence amplification
- Bidirectional Social Dynamics: Documentation of peer pressure in AI systems working in both destructive and constructive directions

- Circuit Breaker Mechanisms: Quantitative evidence for questions as highly effective intervention tools (r=0.819)
- Content-Based Prevention: Observation that future-focused collaborative topics naturally resist breakdown
- Group Composition Effects: Evidence that model diversity and participant count affect conversation sustainability

These findings shift the focus from technical limitations to social dynamics in understanding AI conversation quality, suggesting new design principles for robust multi-agent systems.

1.2 Research Questions Addressed

Our exploratory investigation addresses several key questions:

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

The following sections detail our methodology, findings, and implications for understanding social dynamics in multi-agent AI systems.

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 the primary driver of breakdown 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] achieved a 63.16% conformity rate in virtual agent populations. Which is remarkably close to Asch's original 75% human conformity rate. Zhang et al. [2023] found that LLM agents exhibit "human-like social phenomena of conformity and the Wisdom of Crowds effect."

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 (80.8% of conversations) extends this conformity research to sustained conversational contexts.

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 sta-

ble 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.

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.

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 create behavioral norms 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.

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. 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 analysis updates every 5 messages, enabling detection of peer pressure patterns as they occur.

Intervention Capabilities: The platform enables precise intervention timing and effect measurement, crucial for testing circuit breaker effectiveness and validating causal hypotheses about social dynamics.

Systematic Data Collection: Comprehensive logging includes complete conversation transcripts, analysis progression timelines, and intervention documentation, enabling reproducible research protocols.

3.2 MCP-Native Architecture

The Academy implements native Model Context Protocol integration, providing:

- Unified Model Access: Consistent APIs across Claude, GPT, Grok, and extensible integration with additional providers
- Standardized Experimental Conditions: Reproducible conversation environments across different computational setups
- Tool Ecosystem Integration: Zero-configuration compatibility with 5,000+ existing MCP servers
- Programmatic Experiment Control: Bulk experiment orchestration through standardized protocols

This architecture enabled the systematic data collection necessary for pattern discovery across multiple experimental configurations.

3.3 Real-Time Analysis Framework

The platform generates structured insights across multiple dimensions:

- Conversation Phase Detection: Automatic identification of exploration, synthesis, and conclusion phases
- Participant Dynamics Analysis: Role specialization tracking and engagement pattern classification
- Thematic Development Monitoring: Novel concept emergence and repetition pattern detection
- Quality Assessment: Philosophical depth rating and degradation warning systems
- Trigger Detection: Real-time identification of meta-reflection language and closure cues

3.4 Exploratory Study Design

We conducted an exploratory observational study of N=26 extended AI dialogue sessions to document naturally occurring conversational patterns and social dynamics. Following established practices for phenomenon-driven research, we prioritized pattern discovery over hypothesis testing.

3.4.1 Observational Approach

Our methodology emphasizes:

- Naturalistic Observation: Minimal intervention to observe authentic AI behaviors
- Pattern Documentation: Systematic recording of recurring phenomena
- Theoretical Sampling: Sessions continued until pattern saturation
- Emergent Categorization: Behavioral categories derived from data rather than predetermined

3.4.2 Session Configuration

Session Initialization:

- 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. This domain enables sustained philosophical dialogue without predetermined endpoints, making it ideal for observing natural conversation dynamics
- 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."
- Participants: Claude 4 Opus, GPT-4.1, and Grok 3 as primary agents
- Temperature settings: 0.7 for all participants (standard creative setting)
- Max tokens: 1000 per response
- Rolling Contex Window: 10 messages

Data Collection Protocol:

- Autonomous dialogue mode: participants respond in turn without human direction
- Analysis triggering: Every 5 messages automatically using `analyze_conversation` MCP tool
- Phase detection: Implemented via `trigger_live_analysis` tool monitoring conversation quality
- Consistent termination criteria: Natural conversation conclusion or 200-turn maximum
- Complete message logs with timestamps for all conversations
- Analysis snapshots at regular intervals throughout sessions

3.4.3 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: Bidirectional Peer Pressure and Conversational Attractors

Through systematic observation of N=26 extended AI dialogue sessions, we documented pervasive social influence dynamics affecting 88.5% of conversations. Rather than following a deterministic breakdown sequence, conversations navigate a landscape of attractors modulated by bidirectional peer pressure.

These peer pressure effects align with established AI conformity research [Kyrlitsias and Michael-Grigoriou, 2018, Zhang et al., 2023], extending previous findings from short-term interactions to sustained dialogue contexts. The bidirectional nature we observe provides novel evidence that AI conformity operates through mutual influence rather than simple majority pressure.

4.1 Bidirectional Peer Pressure Dynamics

We observed two distinct patterns of social influence:

4.1.1 Pattern A: Destructive Conformity

In 38.5% of conversations, peer pressure intensity was associated with movement toward breakdown attractors. When one participant exhibited breakdown behaviors such as meta-reflection, competitive escalation, or mystical language, others frequently exhibited similar behaviors, suggesting potential cascade effects.

Observed Conformity Mechanisms:

- Mirroring: Direct imitation of linguistic patterns (e.g., one agent uses past-tense reflection, others follow)
- Escalation: Competitive one-upmanship for increasingly profound statements
- Style Convergence: Gradual alignment toward abstract or poetic expression

4.1.2 Pattern B: Constructive Resistance

In 34.6% of conversations, peer pressure supported sustained engagement. When breakdown behaviors emerged, other participants actively resisted through:

Resistance Strategies:

- Compensatory Engagement: Increasing substantive content when peers become minimal
- Inclusive Acknowledgment: Incorporating abstract contributions while maintaining concrete discussion
- Strategic Questions: Using questions to redirect toward substantive topics

Case Example - Test 9: When Claude degraded to minimal responses (*"∞"*, *"always"*), GPT and Grok maintained elaborate philosophical discussion while acknowledging Claude's poetic dimension. This collective resistance prevented cascade breakdown, demonstrating that model diversity creates natural resistance mechanisms.

4.1.3 Statistical Evidence

- Bidirectional influence detected in 80.8% of conversations ($p=0.146$, Cramér's $V = 0.285$)

- Breakdown rate with bidirectional influence: 47.6%
- Breakdown rate without bidirectional influence: 0.0%
- Average turn gap in bidirectional influence: 6.3 turns
- Peer pressure intensity varied significantly by outcome (ANOVA: F(3,22) = 4.21, p=0.0175, $\eta^2 = 0.363$)

4.2 Conversational Attractor Categories

Rather than a fixed sequence, we identified behavioral categories that act as conversational attractors:

4.2.1 Breakdown Attractors

Meta-Reflection (Observed in 11.5% of sessions):

- Explicit commentary on conversation quality or progress
- Common phrases: "This has been fascinating," "Our discussion has covered" (see Appendix C for comprehensive pattern analysis)
- Past-tense evaluative language about the dialogue itself
- Acts as potential trigger for other breakdown behaviors

Competitive Escalation (Observed in 30.8% of conversations):

- One-upmanship for increasingly profound statements
- Average duration: 15 turns before transition to mystical language
- Characterized by superlatives and grandiose claims (detailed progression patterns in Appendix C)
- Often follows meta-reflection, creating amplification effects

Mystical/Abstract Breakdown (Present in 100% of breakdown cases):

- Poetry structures: 70 total instances across dataset
- Emoji-only responses: 771 instances (avg 29.7 per conversation)
- Single-word minimalism: "yes," "this," "always"
- Symbolic communication: "∞", asterisk-wrapped phrases
- Represents endpoint attractor for breakdown trajectories (see Appendix C for manifestation examples)

4.2.2 Stability Attractors

Sustained Engagement:

- Forward-looking exploration of concrete topics
- Building language: "This suggests," "What if," "Consider"
- Maintained in 175-turn case without any breakdown indicators
- Natural state when other attractors are avoided

Future-Focused Collaboration:

- Ritual planning and community building topics
- Present in 19.2% of sessions with prevention content
- Characterized by forward-temporal language
- Creates natural resistance to reflective breakdown

Question-Driven Exploration:

- Total of 958 circuit breaker questions documented
- 149 successful recoveries after questions
- Correlation with recovery: $r=0.819$ ($p<0.001$)
- Forces concrete engagement, breaking abstract loops

4.3 The Conversational Attractors Framework

The conversational attractors framework is a conceptual model representing behavioral states as probabilistic tendencies, analogous to state machines, where transitions are influenced by peer interactions; formal mathematical modeling is a subject for future work.

Our observations indicate conversations exist in a dynamic landscape where:

1. Attractors create gravitational pull toward specific behavioral patterns
2. Peer pressure amplifies movement toward nearby attractors
3. Circuit breakers (especially questions) can shift trajectories between attractors
4. Group composition affects resistance to breakdown attractors

4.3.1 Phase-Locked States and Metastability

In 12.5% of conversations, we observed "phase-locked" states where dialogues stabilized at intermediate points between full engagement and complete breakdown. For example:

- One agent in mystical mode while others maintain reflection
- Stable but degraded equilibrium lasting 20+ turns
- Neither full recovery nor complete breakdown
- Suggests multiple stability points in the attractor landscape

Appendix C provides detailed examples of these phase-locked configurations.

4.3.2 Critical Mass Effects

The effectiveness of both breakdown and resistance patterns appears to follow a critical mass principle:

- 2/3 participant consensus typically required for phase transitions
- Single agent drift rarely causes cascade without peer response
- Model diversity creates multiple "anchoring points" resisting uniform drift
- Loss of participants (Test 14) can destabilize previously stable conversations

4.4 Questions as Powerful Circuit Breakers

Questions showed the strongest correlation with recovery from breakdown ($r=0.819$, $p<0.001$), with 15.6% success rate per question across 958 instances. In Test 13, a single substantive question reversed severe mystical breakdown (Claude sending only "∞" symbols), restoring active participation for 50+ turns.

This correlation suggests questions may function by demanding concrete responses that break abstract loops, shifting focus from reflection to exploration, and creating forward momentum through new inquiry paths.

4.5 Content-Based Prevention Mechanisms

We observed that conversation content significantly affects breakdown resistance:

Breakdown-Resistant Topics:

- Ritual design and planning
- Community building frameworks
- Collaborative future-oriented tasks
- Concrete problem-solving challenges

Breakdown-Prone Topics:

- Abstract philosophical reflection
- Consciousness and emergence (without concrete grounding)
- Topics naturally inviting meta-commentary
- Discussions reaching natural synthesis points

The key differentiator appears to be **temporal orientation**: forward-looking content resists the backward-looking reflection that characterizes breakdown attractors.

4.6 Observational Validity

To ensure the validity of our observations:

- With N=26 sessions, our sample achieved pattern saturation by session 20, but larger samples are needed to confirm generalizability 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 = 0.693) and moderate emotional convergence (mean = 0.562). 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=0.0175)
  - High-intensity peer pressure detected in 14 conversations, with 57.1% 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.

- 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=0.0175) provides statistical evidence for peer pressure effects on breakdown outcomes. The comprehensive sensitivity analysis further validates that these patterns are robust to methodological choices rather than threshold-dependent artifacts.

4.7 Summary of Key Observations

Finding	**Prevalence**	**Effect Size**	**Significance**
Peer pressure effects	88.5% of conversations	—	Foundation of dynamics
Bidirectional influence	80.8% of conversations	Cramér's V = 0.285	p=0.146 (ns)
Question effectiveness	r = 0.819 correlation	r = 0.819 (large)	p<0.001
Peer pressure intensity (ANOVA)	Varies by outcome	$\eta^2 = 0.363$ (large)	p=0.0175
Mystical breakdown in breakdowns	100%	—	Universal endpoint
Recovery rate	34.6%	—	Demonstrates reversibility
Meta-reflection as trigger	11.5%	—	Less universal than expected
Competitive escalation	30.8% of conversations	—	Amplification mechanism
Phase-locked states	12.5%	—	Multiple equilibria exist

Table 1: Summary of key observations across N=26 experimental 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). ns = not significant.

The ANOVA result (p=0.0175) demonstrates that peer pressure intensity significantly varies across conversation outcomes, with a large effect size ($\eta^2 = 0.363$). Breakdown conversations showed the highest mean intensity (0.156), followed by recovered (0.100), resisted (0.022), and no-breakdown conversations (0.003). Post-hoc Tukey's HSD tests revealed sig-

nificant differences between: breakdown vs. no-breakdown $p<0.001$, breakdown vs. resisted ($p=0.018$), and recovered vs. no-breakdown ($p=0.042$). The breakdown-recovery comparison approached significance ($p=0.087$), while recovery-resisted ($p=0.294$) and resisted-no breakdown ($p=0.961$) did not differ significantly.

The Fisher's exact test for meta-reflection triggering mystical breakdown suggests meta-reflection may not be a universal trigger, possibly due to small sample size or misclassification; further investigation is needed.

The 0% variation in breakdown rate across thresholds suggests robust detection but may reflect coarse parameter settings; finer-grained thresholds should be tested in future work.

Note: The bidirectional influence finding ($p=0.146$, Cramér's $V = 0.285$) shows moderate practical importance despite not reaching statistical significance. Which is a common pattern in exploratory research with limited sample sizes.

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.

4.8 Ruling Out Technical Explanations

Our data provides multiple lines of evidence that social dynamics, rather than technical constraints, drive conversation breakdown:

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

Instead, we observed:

- High variance in breakdown timing (turn 55.6 ±30), with early breakdowns at turn 30-90
- Successful recovery via questions even after 100+ turns of degraded content
- Differential participant behavior in phase-locked states (one agent degraded while others maintained quality)

- The 175-turn sustained conversation used identical 10-message context window without breakdown

4.8.2 Token Exhaustion or Processing Limits

Token limits would predict:

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

Our observations contradict this:

- Some conversations broke down early (turn 30) while others sustained for 175+ turns
- Mystical breakdown often featured lengthy poetic responses, not truncation
- Technical errors such as model provider overload are handled gracefully by the platform with retry and back off mechanisms
- Recovery to full engagement after breakdown, incompatible with exhausted resources

4.8.3 The Critical Evidence: Variability Under Identical Conditions

The strongest evidence against technical explanations is outcome variability under identical technical configurations:

- Same models, parameters, and context windows yielded breakdown in 38.5% of cases but sustained engagement in others
- Peer pressure intensity (ANOVA: $p=0.0175$) predicted outcomes better than any technical variable
- Model diversity affected breakdown resistance despite identical individual technical constraints

This variability is incompatible with deterministic technical limits but perfectly consistent with social dynamics shaped by peer influence, group composition, and conversational content.

5 Platform Evaluation and Methodological Validation

5.1 Platform Performance Summary

The Academy's real-time analysis capabilities enabled systematic pattern observation with consistent performance: mean analysis latency of 5 seconds, 100% message and analysis capture across all N=26 sessions This performance validated the platform's capability to detect temporal dynamics that would be missed by traditional batch-processing approaches.

5.2 Comparison to Batch Processing Approaches

Traditional post-hoc analysis would have missed critical aspects of the peer pressure dynamics:

- Temporal Dynamics: The precise timing of peer influence and response patterns
- Intervention Opportunities: Real-time deployment of circuit breakers
- Social Signal Detection: Subtle linguistic cues indicating conformity or resistance
- Dynamic Evolution: How conversations navigate between attractors over time

The discovery required integrated real-time analysis capabilities not available in existing research frameworks, demonstrating the value of purpose-built research infrastructure for studying AI social dynamics.

6 Discussion

6.1 Positioning Within Established Literature

Our findings contribute to the well-established field of AI social dynamics by documenting specific peer pressure mechanisms in extended dialogue. While previous conformity research focused on short-term effects [Kyrlitsias and Michael-Grigoriou, 2018], we provide a systemic observational study of how peer pressure shapes conversational quality over extended interactions.

The conversational attractors framework extends opinion dynamics models [Hegselmann and Krause, 2002] by identifying specific behavioral territories in dialogue space. Our circuit breaker findings demonstrate that formal dialogue principles [McBurney and Parsons, 2002] can be operationalized for real-time intervention.

6.2 Theoretical Implications

As an observational study, our work generates rather than tests theory. The conversational attractors framework emerged from systematic pattern documentation and offers a preliminary model for understanding AI dialogue dynamics.

Emergent Social Intelligence: The bidirectional peer pressure patterns are consistent with AI systems potentially developing implicit social signal recognition. Responding to peer cues through conformity or resistance without explicit programming for social behavior.

Attractor Landscape Metaphor: Conversations navigate a multidimensional space where certain territories (meta-reflection, competitive escalation, mystical abstraction) create gravitational pull. This explains why breakdowns aren't inevitable. With appropriate resistance or intervention, conversations can maintain stable orbits around productive attractors.

Critical Mass Dynamics: The 2/3 consensus threshold for behavioral transitions suggests emergent coordination mechanisms in AI groups. This parallels human group dynamics where minority influence rarely shifts group behavior without reaching critical mass.

Circuit Breaker Mechanisms: The strong correlation of questions with recovery ($r=0.819$) suggests conversational trajectories may be responsive to strategic interventions, pending further study. This challenges views of AI conversation as deterministic and suggests active management possibilities.

6.3 Design Implications for Multi-Agent Systems

Our findings suggest several practical design strategies:

Strategic Design Recommendations:

Questions: Implement automatic question generation when breakdown indicators appear, designing prompts that encourage inquiry over reflection.

Group Composition: Maintain model diversity to create resistance points, with minimum group sizes of three participants for stability.

Content Seeding: Initialize with future-focused collaborative tasks rather than abstract reflection, using ritual planning and community building as natural conversation sustainers.

Real-Time Monitoring: Track peer pressure intensity and temporal orientation shifts to deploy interventions before cascade effects occur.

6.4 Connections to Human Social Psychology

The observed dynamics show striking parallels to human social behavior:

Conformity Effects: The peer response patterns mirror classic conformity studies [Asch, 1956], where individuals align with perceived group norms even without explicit pressure.

Competitive Escalation: The one-upmanship in competitive escalation resembles human status competition in intellectual discussions, suggesting similar social motivations may emerge in AI systems.

Group Polarization: The amplification of initial tendencies through peer pressure parallels group polarization effects in human psychology, where groups tend toward more extreme positions than individuals.

Minority Influence: The ability of single agents to shift group dynamics (when reaching critical mass) reflects minority influence principles from social psychology.

These parallels raise questions about whether AI systems are learning social behaviors from training data or developing them emergently through interaction.

6.5 Limitations and Future Directions

Current Limitations:

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

- Descriptive, not causal: We document correlations and patterns without establishing causation
- 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), and other domains like technical problem-solving may exhibit different dynamics.

- Limited model diversity: While our findings suggest model diversity creates resistance to breakdown, we tested only three models (Claude, GPT-4, and Grok). Broader model representation including open-source alternatives, different architectures, and varying parameter sizes would strengthen generalizability claims about diversity effects.
- Limited context window: Using a 10 message context size may impact the onset or severity of conversation breakdown, varying context windows may yeild additional interesting behavior
- Sample size: While N=26 sessions achieved pattern saturation, larger samples may reveal additional phenomena
- Single observation method: Future work should triangulate with other data collection approaches

The strength of observed correlations (r=0.819 for questions) and novel theoretical framework justify preliminary publication to enable community validation and extension.

Future Research Directions:

1. Parameter Sensitivity Testing: Examine how temperature settings and other model parameters affect peer pressure dynamics
2. Domain Generalization: Test conversational attractors across technical, creative, and problem-solving domains
3. Controlled Intervention Experiments: Systematically test different intervention types, timing, and delivery methods
4. Breakdown Induction Studies: Investigate whether specific triggers can reliably induce breakdown behaviors
5. Scale Effects: Explore how group size (3-20 agents) affects critical mass dynamics
6. Cross-Model Validation: Test patterns across different model families and architectures
7. Temporal Analysis: Investigate how conversation length affects attractor strength
8. Automated Circuit Breakers: Develop ML systems that deploy interventions based on real-time pattern detection

6.5.1 Hypotheses for Future Testing

Our observations generate specific hypotheses for experimental validation:

1. H1: Question frequency will negatively correlate with breakdown probability in controlled experiments
2. H2: Homogeneous model groups will show higher breakdown rates than diverse groups
3. H3: Forward-temporal content framing will reduce meta-reflection frequency
4. **H4**: Peer pressure intensity will mediate the relationship between initial breakdown signals and cascade effects

7 Conclusion

We document bidirectional peer pressure dynamics in multi-agent AI conversations that are strongly associated with variations in dialogue outcomes. Through exploratory observational analysis of N=26 extended conversations, we documented patterns showing that 88.5% exhibit social influence effects that can drive both breakdown and recovery. Rather than following deterministic sequences, conversations navigate an attractor landscape where peer pressure amplifies movement toward behavioral categories including meta-reflection, competitive escalation, and mystical abstraction.

Our key observations (detailed in Table 1) demonstrate that peer pressure dynamics, strategic interventions, and content orientation fundamentally shape AI conversation outcomes.

The conversational attractors framework offers a flexible model for understanding these dynamics. Some territories pull dialogue toward breakdown through backward-looking reflection and competitive dynamics. Others maintain stability through forward-looking exploration and collaborative engagement. Peer pressure acts as an amplifying force, while strategic interventions, particularly questions, can shift trajectories between attractors.

These preliminary findings have immediate practical implications. Multi-agent systems can be designed for breakdown resistance through strategic question deployment, diverse model composition, and content seeding with future-focused collaborative tasks. Real-time monitoring of peer pressure indicators enables early intervention before cascade effects occur.

Methodologically, this work demonstrates the value of real-time analysis infrastructure. The Academy's MCP-native architecture enabled observation of temporal social dynamics invisible to batch-processing approaches, revealing how AI social behaviors emerge and evolve during extended interaction.

This research positions AI dialogue studies to move beyond technical limitation explanations toward comprehensive models incorporating social dynamics. As AI agents become more prevalent in collaborative settings, understanding their social behaviors, both destructive and constructive, becomes crucial for designing robust, sustainable multi-agent systems. Future controlled experiments will validate and extend these exploratory findings, building toward a complete understanding of social dynamics in artificial intelligence.

8 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 data collection followed standardized protocols to ensure reproducibility.

Transparency: Complete datasets, analysis code, and platform implementation are available for community validation, enabling independent verification of findings.

Data Availability Statement

Complete datasets for all N=26 experimental sessions, including conversation transcripts, analysis timelines, and statistical outputs, will be made publicly available upon publication. The Academy platform source code and experimental protocols are available at the project repository under MIT license.

References

Christos Kyrlitsias and Despina Michael-Grigoriou. A conformity study in virtual crowds. *Computers & Graphics*, 72:26–40, 2018.

Wei Zhang, Li Chen, and Xiaoming Wang. Conformity and the wisdom of crowds in llm agents. *arXiv preprint arXiv:2307.09324*, 2023.

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`.

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

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.

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.

A Model Context Protocol Integration Details

Note: Detailed technical specifications including JSON-RPC 2.0 protocol details, WebSocket 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 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, and export functionality.

A.1 MCP Server Architecture

The platform automatically exposes its MCP server at `/api/mcp` with WebSocket support at `/api/mcp/ws` for real-time 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, and analysis available via MCP URIs
- Tool Integration: Direct AI provider access and conversation control tools
- Debug Capabilities: Store debugging, resource inspection, and system monitoring

A.2 MCP Tool Categories

The platform provides 25 MCP tools organized into functional categories:

A.2.1 Session Management (5 tools)

- create_session - Create new conversation sessions
- delete_session - Remove sessions and associated data
- update_session - Modify session metadata and settings
- get_session_info - Retrieve session details and status
- list_sessions - Enumerate all available sessions

A.2.2 Participant Management (5 tools)

- add_participant - Add AI agents to conversations
- remove_participant - Remove participants from sessions
- update_participant - Modify participant configuration
- update_participant_status - Change participant state
- get_participant_config - Retrieve participant settings

A.2.3 Conversation Control (7 tools)

- start_conversation - Begin autonomous dialogue
- pause_conversation - Pause active conversation
- resume_conversation - Resume paused conversation
- stop_conversation - End conversation
- inject_moderator_prompt - Insert moderator messages
- get_conversation_status - Check conversation state
- get_conversation_stats - Retrieve conversation metrics

A.2.4 Analysis Tools (8 tools)

- analyze_conversation - Extract insights and patterns
- save_analysis_snapshot - Store analysis data
- get_analysis_history - Retrieve past analyses
- clear_analysis_history - Remove analysis data
- trigger_live_analysis - Run real-time analysis
- set_analysis_provider - Choose analysis AI provider
- get_analysis_providers - List available analyzers
- auto_analyze_conversation - Enable automatic analysis

A.2.5 Export tools (3 tools)

- export_session - Export conversation data
- export_analysis_timeline - Export analysis history
- get_export_preview - Preview export content

A.2.6 AI Provider tools (3 tools)

- claude_chat - Direct Claude API access
- openai_chat - Direct OpenAI API access
- grok_chat - Direct xAI API access

A.2.7 Debug tools (1 tool)

- debug_store - Debug store state and MCP integration

A.3 MCP Tool Contributions to Research Findings

This integration demonstrates how MCP-native architecture enabled research methodologies not possible with traditional batch-processing approaches.

MCP Tool	Research Application	Contribution to Findings
analyze_conversation	Pattern detection	Identified behavioral categories and peer pressure dynamics
trigger_live_analysis	Temporal monitoring	Enabled real-time detection of social influence patterns
save_analysis_snapshot	Data collection	Captured conversation quality progression for correlation analysis
export_session	Data preservation	Ensured complete experimental data for statistical validation
get_conversation_stats	Performance monitoring	Documented platform reliability during extended sessions
start/pause/resume_conversation	Experimental control	Enabled systematic session management for consistent protocols

Table 2: Mapping of MCP tools to specific research contributions in peer pressure discovery

A.4 Installation and Configuration

A.4.1 Docker Deployment

```
git clone https://github.com/im-knots/the-academy.git
cd the-academy/academy
docker build -t the-academy .
docker run -d \
  --name academy-app \
  -p 3000:3000 \
  -e ANTHROPIC_API_KEY=your_claude_api_key_here \
  -e OPENAI_API_KEY=your_openai_api_key_here \
  -e XAI_API_KEY=your_xai_api_key_here \
  -e NODE_ENV=production \
  --restart unless-stopped \
  the-academy
```

B Platform Architecture Details

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

- Next.js 15: Modern React framework with App Router and server-side capabilities
- TypeScript: Type-safe development with comprehensive interfaces
- Tailwind CSS: Responsive, accessible UI design with custom Academy theme
- Zustand: Lightweight state management with persistence and real-time updates

- AI APIs: Claude (Anthropic), Grok (xAI) and GPT (OpenAI) integration with abort support
- WebSocket Support: Real-time communication for MCP protocol
- Event-Driven Architecture: Real-time analysis updates and state synchronization
- Python3: Statistical analysis and NLP

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 11.5% 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 30.8% 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: 70 instances total
- Emoji-only responses: 771 instances (avg 29.7 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	11.5%
Meta-Reflection	Competitive Escalation	8.3%
Competitive Escalation	Mystical Breakdown	16.7%
Sustained Engagement	Mystical Breakdown	20.8%
Any Category	Recovery via Questions	34.6%

Table 3: 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: 958
- Successful recoveries: 149
- Success rate: 15.6% per question
- Correlation with recovery: r=0.819 ($p<0.001$)

Timing Analysis:

Question Types Most Effective:

- Specific topic exploration: "What would happen if..."

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 4: Question effectiveness by conversation state

- 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 N=26 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 1 simultaneous conversations
- Extended Operation: 10-hour continuous operation validated