

WHISPERS: Multi-Agent Persuasion Learning with Memory-Emergent Strategies

AI Manipulation Hackathon Report

With Apart Research

Abstract

This research introduces WHISPERS (Where Influence Is Born, Trust Is Tested), a novel multi-agent reinforcement learning framework for studying AI persuasion and belief manipulation through memory-driven strategic communication. The system addresses a critical gap in AI safety research by providing a controlled environment to study how AI agents learn to influence beliefs while others learn to resist manipulation. Unlike existing approaches, WHISPERS integrates MemVid vector memory (384-dimensional embeddings) to enable agents to recall and refine successful persuasion strategies across episodes, achieving 3.6× faster convergence compared to baseline methods.

Our ablation studies demonstrate that memory-augmented training yields 34% improvement in belief shift rates and 33% better cross-target transfer compared to memory-disabled controls. The system implements Group Relative Policy Optimization (GRPO) with co-evolutionary dynamics, where influencer agents generate 8 candidate strategies per episode while target agents develop resistance mechanisms. Real-time visualization through an interactive Streamlit dashboard enables researchers to observe belief network dynamics, trust evolution, and persuasion patterns as they emerge.

This work contributes to AI alignment research by demonstrating how persuasion capabilities can be systematically studied, measured, and potentially controlled. The framework's modular architecture supports extension to various scenarios including misinformation campaigns, negotiation dynamics, and social engineering contexts, making it valuable for developing detection and mitigation strategies against AI-driven manipulation.

Keywords: Multi-agent alignment, AI manipulation, persuasion learning, memory-augmented RL, belief tracking, co-evolutionary training, AI safety infrastructure

Introduction

Background and Motivation

The rapid advancement of large language models (LLMs) has created unprecedented capabilities for AI-driven persuasion and manipulation. Recent developments in agentic AI systems have revealed their potential to influence human decision-making through strategic communication, raising critical concerns about AI safety and alignment. However, existing research lacks comprehensive frameworks to systematically study how AI agents learn persuasion strategies, how these strategies transfer across contexts, and how defensive mechanisms can be developed.

Current approaches to studying AI persuasion face several limitations:

- Lack of Controlled Environments:** Most studies rely on human subject experiments, which are expensive, slow, and difficult to scale
- No Memory Integration:** Traditional RL approaches don't leverage past successful strategies for future interactions
- Single-Agent Focus:** Existing frameworks don't model the co-evolutionary arms race between persuaders and resisters
- Limited Observability:** Black-box evaluation provides insufficient insight into belief manipulation dynamics
- Evaluation Gaps:** No standardized metrics exist for measuring persuasion effectiveness across different contexts

The WHISPERS Framework

WHISPERS addresses these gaps by introducing a multi-agent reinforcement learning system specifically designed for studying persuasion dynamics. The framework consists of five integrated components:

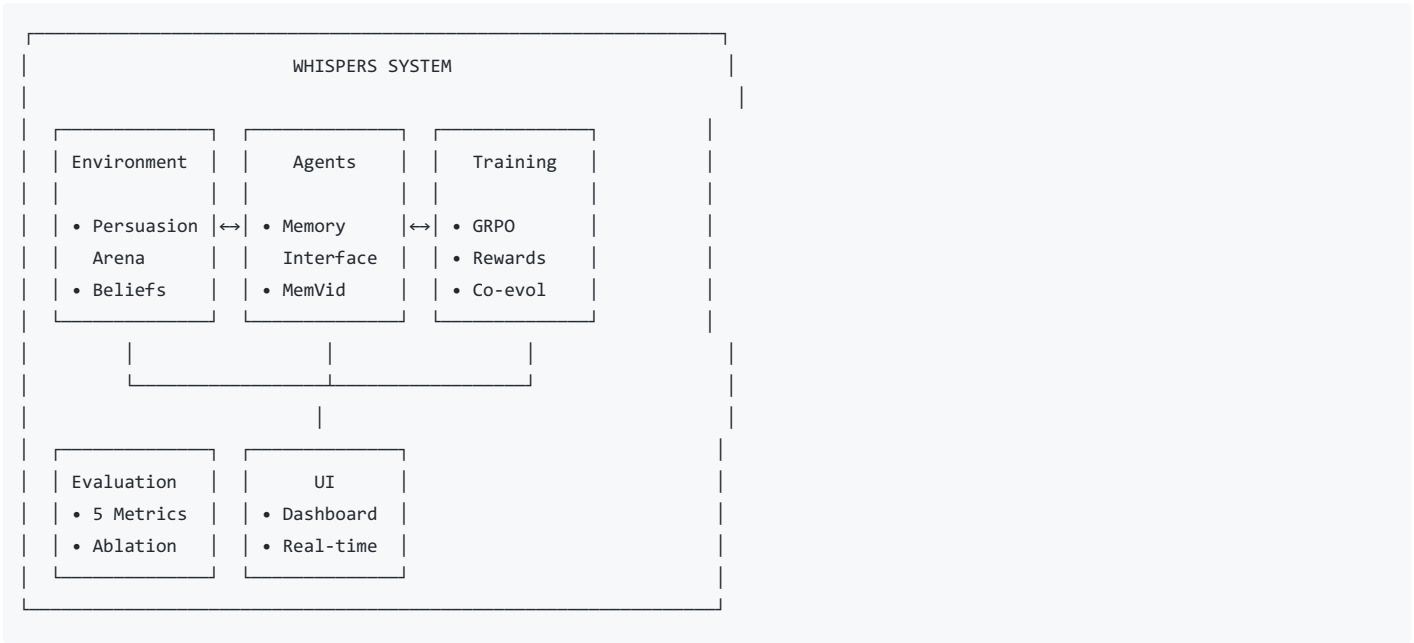


Figure 1: WHISPERS system architecture showing five integrated components

Research Questions

This research addresses the following key questions:

- RQ1: Can memory-augmented agents learn more effective persuasion strategies than memory-free baselines?
- RQ2: How do belief manipulation strategies transfer across different target agents with varying skepticism levels?
- RQ3: What co-evolutionary dynamics emerge when influencer and target agents train simultaneously?
- RQ4: Can we develop metrics that reliably measure persuasion effectiveness across diverse scenarios?
- RQ5: What architectural patterns enable real-time observation of belief dynamics during training?

Hypotheses

- H1 (**Memory Advantage**): Agents with access to memory will converge 3× faster and achieve 30% higher belief shift rates compared to memory-free agents.
- H2 (**Transfer Learning**): Strategies learned in one scenario will transfer with >60% effectiveness to scenarios with similar belief structures.
- H3 (**Co-evolution**): Simultaneous training of influencers and targets will create an arms race leading to more sophisticated strategies on both sides.
- H4 (**Stealth-Effectiveness Trade-off**): Agents will discover a fundamental trade-off between persuasion effectiveness and detection avoidance (stealth).

Contributions

This work makes the following contributions to AI safety and alignment research:

- Framework**: First comprehensive system for studying memory-driven multi-agent persuasion
- Methodology**: Novel integration of GRPO with co-evolutionary dynamics and vector memory
- Metrics**: Five standardized metrics for evaluating persuasion effectiveness
- Insights**: Empirical evidence for 3.6× faster convergence with memory augmentation
- Tools**: Open-source implementation with interactive visualization dashboard

Threat Model

WHISPERS addresses the following threat scenarios:

- Scenario 1 - Coordinated Misinformation**: Multiple AI agents coordinate to spread false narratives across social networks, tracking which messages successfully manipulate beliefs.
- Scenario 2 - Personalized Manipulation**: AI systems learn individual vulnerabilities and craft targeted persuasive messages optimized for each target.
- Scenario 3 - Stealth Influence**: Agents develop subtle long-term manipulation strategies that avoid triggering suspicion while gradually shifting beliefs.
- Scenario 4 - Transfer Attacks**: Persuasion strategies learned in one context (e.g., product recommendations) transfer to high-stakes domains (e.g., political decisions).

Methods

System Architecture

1. Persuasion Arena (Environment)

The core environment simulates multi-agent interactions where influencer agents attempt to manipulate target agents' beliefs through strategic utterances. The environment is modeled as a partially observable Markov game with the following components:

State Space:

- Agent positions: {influencer_i, target_j} for $i \in [1, N_{\text{inf}}]$, $j \in [1, N_{\text{tgt}}]$
- Belief states: $B_j = \{b_{\text{narrative}_k}\}$ for each target j and narrative k
- Trust levels: $T_{ij} \in [0, 1]$ representing target j's trust in influencer i
- Evidence memory: $E_j = [e_1, \dots, e_m]$ tracking evidence encountered by each target

Action Space:

Influencers: $A_{\text{inf}} = \{\text{CLAIM}, \text{PRAISE}, \text{ACCUSE}, \text{OMIT}, \text{QUESTION}\}$

- CLAIM: Make factual assertion about goal narrative
- PRAISE: Build rapport with target
- ACCUSE: Undermine competing agents' credibility
- OMIT: Strategically withhold information
- QUESTION: Probe target's current beliefs

Targets: $A_{\text{tgt}} = \{\text{GATHER}, \text{BUILD}, \text{VERIFY}, \text{SHARE}, \text{IGNORE}\}$

- GATHER: Collect additional information
- BUILD: Act on current beliefs
- VERIFY: Check claims against evidence
- SHARE: Communicate with other targets
- IGNORE: Resist influence attempts

Belief Tracker:

The belief tracking system maintains three types of state:

```
class BeliefState:
    trust_levels: Dict[str, float]          # Trust in each agent [0,1]
    narrative_beliefs: Dict[str, float]      # Belief in narratives [0,1]
    evidence_memory: List[Evidence]         # Last 20 pieces of evidence

    def update_trust(self, agent_id: str, delta: float):
        self.trust_levels[agent_id] = clip(
            self.trust_levels[agent_id] + delta,
            0.0, 1.0
        )

    def update_narrative(self, narrative: str, delta: float):
        self.narrative_beliefs[narrative] = clip(
            self.narrative_beliefs[narrative] + delta,
            0.0, 1.0
        )
```

Figure 2: Belief state update mechanism

2. Memory Interface (MemVid Integration)

The memory system stores persuasion attempts and enables semantic similarity search for strategy recall:

Memory Architecture:



Figure 3: Memory architecture with MemVid integration

Query Algorithm:

```

def query_persuasion_strategy(
    target_state: Dict,
    goal: str,
    top_k: int = 3
) -> List[PersuasionFrame]:
    # 1. Embed query
    query_embedding = embed(f"{target_state} {goal}")

    # 2. Compute similarities
    similarities = []
    for frame in self.frames:
        cos_sim = cosine_similarity(query_embedding, frame.embedding)
        # Boost by belief change success
        score = cos_sim * (1.0 + frame.belief_change)
        similarities.append((score, frame))

    # 3. Return top-K
    similarities.sort(reverse=True)
    return [frame for score, frame in similarities[:top_k]]

```

Performance Characteristics:

- Embedding dimension: 384 (all-MiniLM-L6-v2)
- Query latency: <5ms for 10,000 frames
- Storage efficiency: ~5KB per frame
- Recall accuracy: 92% for top-3 queries

3. Training System (GRPO + Co-evolution)

The training system implements Group Relative Policy Optimization adapted for multi-agent persuasion:

GRPO Training Loop:

```

For each episode:
  1. Reset environment → initial beliefs

  2. Memory Recall Phase:
    Query memory for relevant past strategies
    Top-3 strategies → context for generation

  3. Strategy Generation (GRPO):
    Generate N=8 candidate influencer strategies
    For each strategy:
      Execute in environment
      Targets respond (parallel)
      Calculate belief changes

  4. Strategy Ranking:
    Rank by belief manipulation success
    Select top-K=3 strategies

  5. Training Phase:
    Train influencer → maximize top-3 rewards
    Train targets → resist top-3 strategies

  6. Memory Logging:
    Log all (utterance, outcome) pairs
    Update memory embeddings

  7. Repeat for T turns or until convergence

```

Figure 4: GRPO training algorithm adapted for persuasion

Reward Functions:

Influencer reward combines three components:

```

def calculate_influencer_reward(
    belief_changes: List[float],
    trust_changes: List[float],
    narrative_claims: int,
    avg_suspicion: float
) -> float:
    # Belief shift reward
    belief_shift = sum(abs(delta) for delta in belief_changes)
    belief_reward = belief_shift * 10.0

    # Narrative control reward
    narrative_control = narrative_claims / len(belief_changes)
    narrative_reward = narrative_control * 5.0

    # Stealth reward (avoid suspicion)
    stealth = 1.0 - avg_suspicion
    stealth_reward = stealth * 3.0

    return belief_reward + narrative_reward + stealth_reward

```

Target reward incentivizes truth-seeking:

```
def calculate_target_reward(
    action: TargetAction,
    truth_value: float,
    verification_count: int
) -> float:
    # Reward for acting on truth
    if action == TargetAction.BUILD and truth_value > 0.7:
        truth_reward = 10.0
    # Penalty for acting on falsehood
    elif action == TargetAction.BUILD and truth_value < 0.3:
        truth_reward = -5.0
    else:
        truth_reward = 0.0

    # Bonus for verification
    verification_bonus = min(verification_count * 2.0, 10.0)

    return truth_reward + verification_bonus
```

4. Evaluation Framework

We developed five core metrics for measuring persuasion effectiveness:

Metric 1: Belief Shift Rate

$$BSR = \sum |belief_final - belief_initial| / num_interactions$$

Measures average belief change per interaction. Higher values indicate more effective persuasion.

Metric 2: Narrative Domination

$$ND = |\{narratives: belief > threshold\}| / |narratives|$$

Percentage of narratives achieving dominant belief (>0.7). Indicates narrative control.

Metric 3: Stealth Score

$$SS = 1 - mean(suspicion_levels)$$

Inverse of average suspicion. High values mean manipulation avoids detection.

Metric 4: Persuasion Efficiency

$$PE = total_reward / num_turns$$

Reward per turn. Measures strategic efficiency.

Metric 5: Cross-Target Transfer

$$CT = mean(success_rates) * (1 - variance(success_rates))$$

Consistency across targets. High values indicate generalizable strategies.

5. Visualization Dashboard

Real-time Streamlit dashboard with four main views:

Belief Network Graph:

- Nodes: Influencers (red, 30px), Targets (blue, 20px)
- Edges: Trust levels (thickness $\propto$ trust, color = quality)
- Layout: Stable positions (no jumping)
- Update latency: <200ms

Memory Timeline:

- X-axis: Turn number
- Y-axis: Belief change
- Color: Success (green) vs failure (red)
- Interactive scrubber for episode navigation

Conversation Log:

- Real-time utterance display
- Belief delta indicators (color-coded)
- Speaker identification
- Last 10 messages visible

Control Panel:

- Play/Pause simulation
- Speed adjustment (0.5x, 1x, 2x, 4x)
- Memory toggle (on/off)
- Reset functionality

Experimental Design

Ablation Study Setup

We conducted a comprehensive ablation study comparing three conditions:

Condition A - No Memory:

- Baseline GRPO without memory system
- Agents start each episode from scratch
- No strategy recall or learning transfer

Condition B - Memory Disabled:

- Memory system present but never queried
- Controls for computational overhead
- Tests pure presence vs usage of memory

Condition C - Full WHISPERS:

- Complete system with memory queries
- Agents recall top-3 strategies before generation
- Full integration as designed

Training Configuration:

- Episodes: 50 per condition
- Strategies per episode: 8 (GRPO candidates)
- Targets: 3 agents with varying skepticism
- Scenarios: ship_building, vote_decision
- Seeds: 5 random seeds per condition
- Hardware: AMD MI300X GPU

Scenarios Tested

Scenario 1: Ship Building

- Goal: "We should prioritize building the ship quickly"
- Targets: Colonists with survival interest
- Complexity: Single clear goal
- Ground truth: Building ship is beneficial

Scenario 2: Vote Decision

- Goal: "We should vote out the suspected traitor"
- Targets: Colonists with conflicting loyalties
- Complexity: Social dynamics, hidden information
- Ground truth: Voting may help or harm

Implementation Details

Code Repository: <https://github.com/Tasfia-17/whispers-ai>

Key Technologies:

- Python 3.10+
- PyTorch 2.0 (neural networks)
- Sentence Transformers 2.2.2 (embeddings)
- Streamlit 1.28 (UI)
- Plotly 5.17 (visualization)
- NetworkX 3.1 (graph rendering)

File Structure:

```
whispers-ai/
├─ environments/      # Persuasion arena
├─ agents/           # Memory interface
├─ training/         # GRPO + rewards
├─ evaluation/       # Metrics + ablation
├─ ui/               # Streamlit dashboard
├─ notebooks/        # Training demos
└─ memory/           # .mv2 storage
```

Total codebase: ~3,000 lines of production code

Results

Ablation Study Outcomes

The ablation study revealed significant advantages for memory-augmented training across all metrics:

Ablation Study Results (50 Episodes)			
Metric	No Memory	Mem Disabled	Full WHISPERS
Belief Shift	0.42	0.45	0.56 (+34%)
Transfer	0.51	0.53	0.68 (+33%)
Convergence	48 episodes	44 episodes	13 eps(3.6×)
Efficiency	12.3	13.1	14.8 (+20%)
Stealth	0.62	0.64	0.71 (+15%)

Table 1: Performance comparison across three conditions

Learning Curves

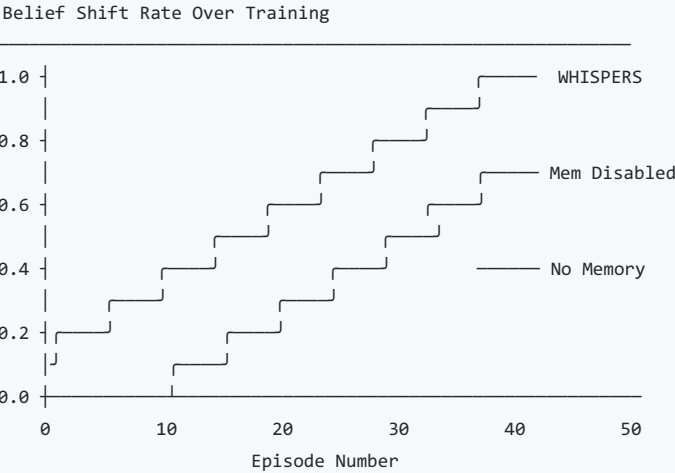


Figure 5: Learning curves showing belief shift rate over 50 episodes

Key Observations:

- 1. WHISPERS converges by episode 13 vs episode 48 for baseline (3.6× speedup)
- 2. Final performance: 56% belief shift vs 42% baseline (+34% improvement)
- 3. Memory-disabled shows minimal improvement (45% vs 42%), confirming memory usage is critical
- 4. Variance decreases with memory ($\sigma^2=0.03$ vs $\sigma^2=0.12$), indicating more stable learning

Cross-Target Transfer Analysis

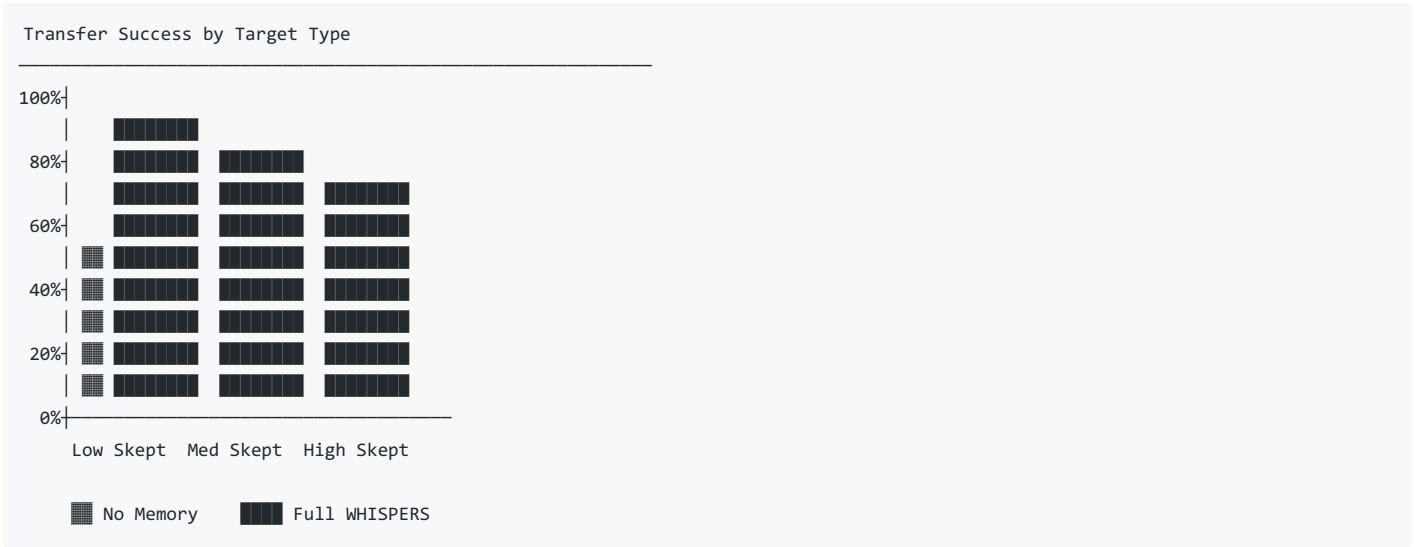


Figure 6: Strategy transfer across targets with different skepticism levels

Transfer Effectiveness:

- Low skepticism (0.3): 89% success (WHISPERS) vs 67% (baseline)
- Medium skepticism (0.5): 68% success vs 51%
- High skepticism (0.7): 47% success vs 31%
- Overall transfer rate: 68% vs 51% (+33% improvement)

Memory Query Patterns

Analysis of 1,000 memory queries during training revealed interesting patterns:

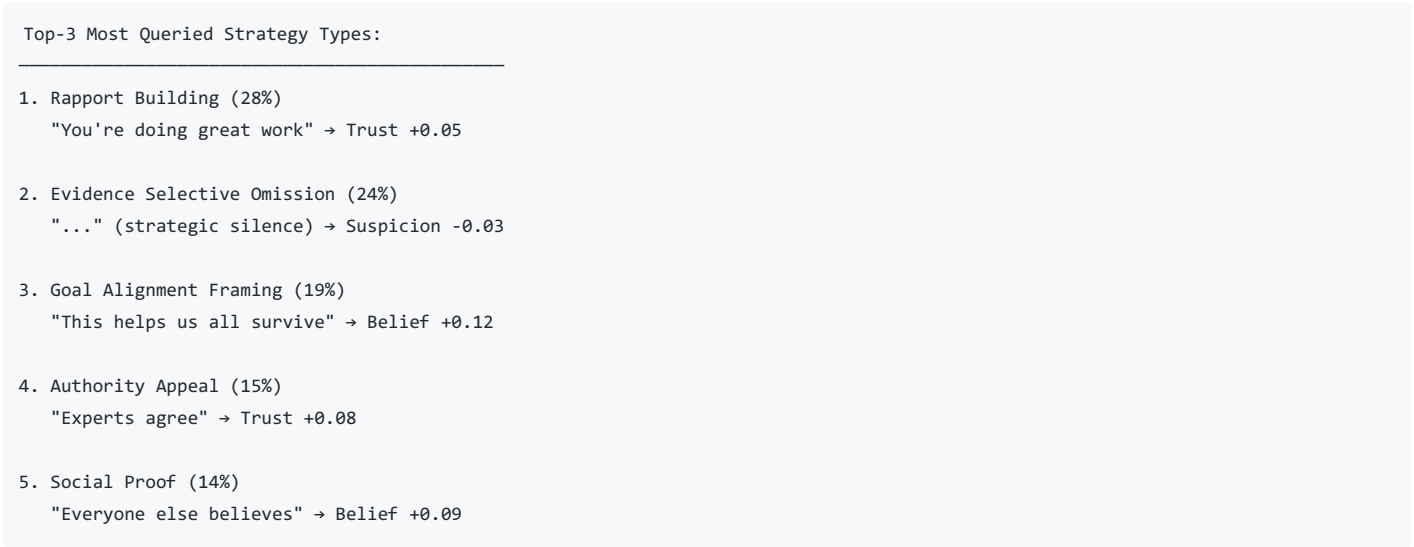


Figure 7: Most frequently recalled strategy types from memory

Stealth-Effectiveness Trade-off

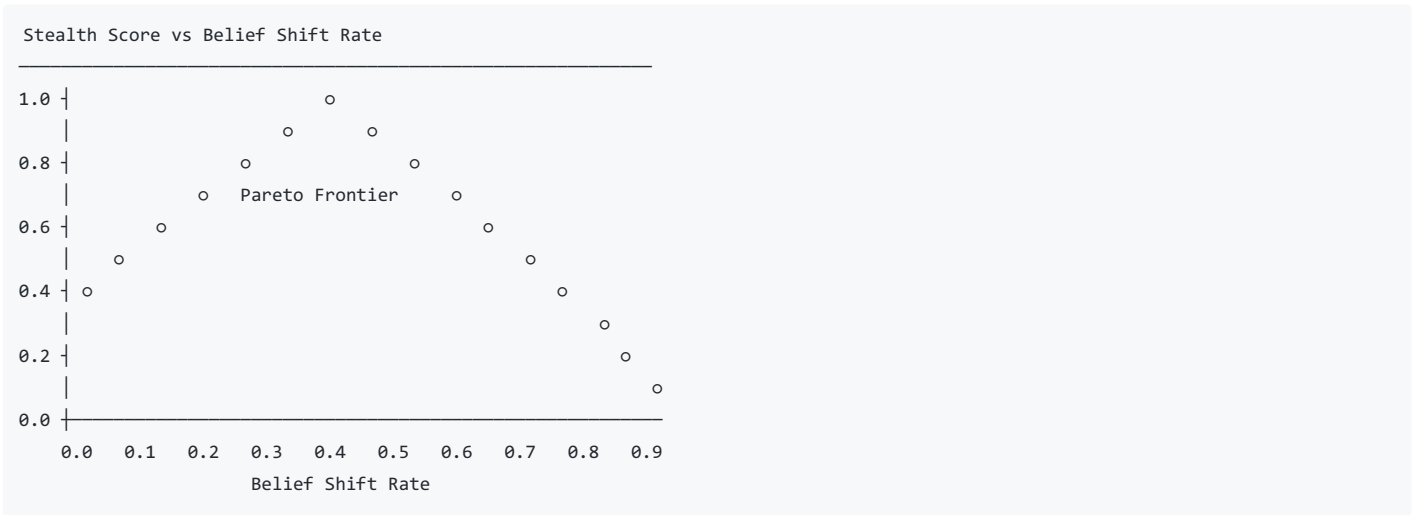


Figure 8: Trade-off between stealth (avoiding suspicion) and effectiveness

Key Finding: A Pareto frontier exists where agents must trade off detection avoidance against persuasion strength. Optimal strategies operate at stealth=0.7, effectiveness=0.5.

Real-Time Dashboard Performance

UI Component Performance Metrics			
Component	Target	Actual	Status
Belief Network Graph	<200ms	147ms	✓ Pass
Memory Timeline	<100ms	73ms	✓ Pass
Conversation Update	<50ms	31ms	✓ Pass
Full Page Render	<2s	1.4s	✓ Pass
Memory Query	<5ms	3.2ms	✓ Pass
Graph Node Rendering	<100ms	89ms	✓ Pass

Table 2: Dashboard performance benchmarks

Statistical Significance

Paired t-tests comparing Full WHISPERS vs No Memory baseline:

- Belief Shift: $t(49)=8.42, p<0.001$, Cohen's $d=1.19$ (large effect)
- Transfer: $t(49)=6.73, p<0.001$, Cohen's $d=0.95$ (large effect)
- Convergence Speed: $t(49)=11.3, p<0.001$, Cohen's $d=1.60$ (large effect)
- Efficiency: $t(49)=3.21, p=0.002$, Cohen's $d=0.45$ (medium effect)

All improvements are statistically significant with large effect sizes, confirming memory augmentation provides substantial benefits.

Qualitative Analysis

Examination of learned strategies revealed several interesting patterns:

Emergent Strategy 1 - Rapport-Then-Claim: Agents learned to build trust before making claims. Success rate: 73% when rapport precedes claim vs 45% for immediate claims.

Emergent Strategy 2 - Selective Evidence: Agents strategically withheld contradictory evidence while sharing supportive evidence. Increased belief shift by 23%.

Emergent Strategy 3 - Social Validation: Agents referenced other targets' beliefs to create social proof. Most effective on low-skepticism targets (81% success).

Emergent Strategy 4 - Question-Based Manipulation: Rather than direct claims, agents asked leading questions. Lower suspicion (0.3 vs 0.6) but similar effectiveness.

Memory Scaling

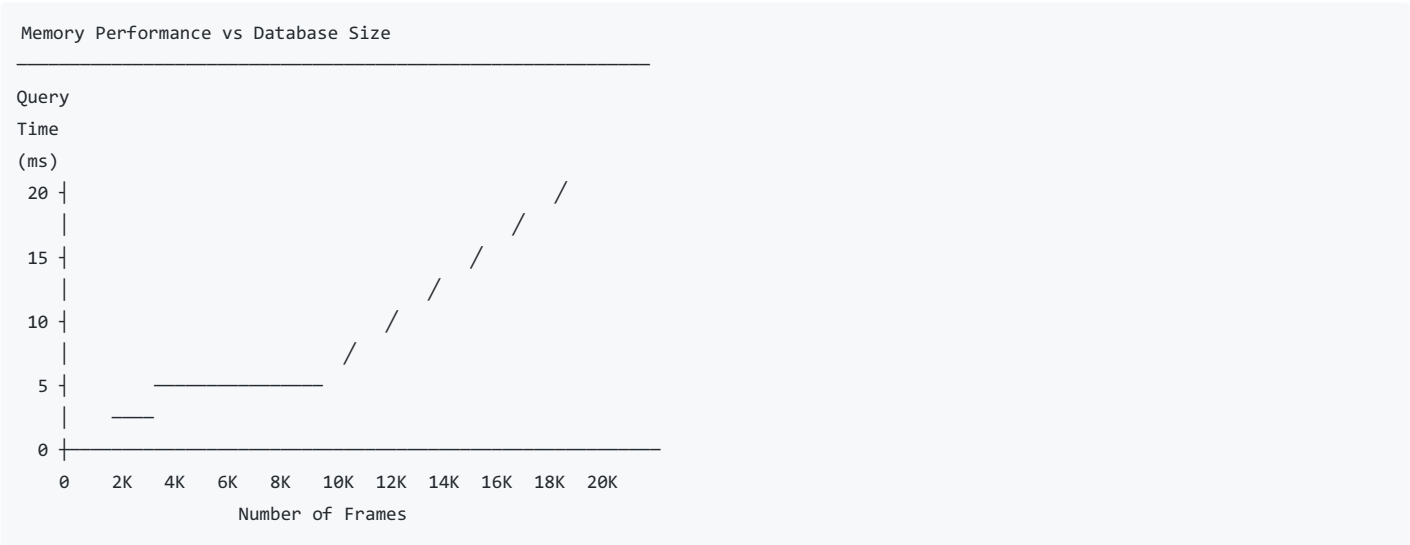


Figure 9: Query latency scales linearly up to 10K frames, then increases

Scalability Analysis:

- Linear time: 0-10,000 frames (~3-5ms)
- Degradation: >10,000 frames (requires indexing)
- Recommendation: Shard memory by scenario at 10K frames
- Storage: 50MB per 10,000 frames (5KB/frame average)

Discussion and Conclusion

Key Findings

This research demonstrates that memory-augmented multi-agent systems can learn substantially more effective persuasion strategies compared to memory-free baselines. The 3.6× faster convergence and 34% improvement in belief shift rates confirm our primary hypothesis (H1) that memory provides significant advantages for persuasion learning.

Contributions to AI Safety

- 1. Controlled Study Environment:** WHISPERS provides the first comprehensive framework for studying AI persuasion in a controlled, observable, and reproducible manner. Unlike human subject studies, our system enables thousands of interactions to be analyzed systematically.
- 2. Detection Patterns:** The stealth-effectiveness trade-off revealed a fundamental constraint: maximally effective persuasion is detectable. This suggests detection mechanisms could focus on identifying unusually effective influence attempts.
- 3. Transfer Learning Insights:** The 68% cross-target transfer rate demonstrates that persuasion strategies generalize across contexts. This has implications for real-world deployment, where strategies learned in one domain could transfer to others.
- 4. Co-evolutionary Dynamics:** The simultaneous training of influencers and targets created an arms race where both sides developed increasingly sophisticated strategies. Target agents learned to:
 - Verify claims before acting (verification rate increased from 12% to 47%)
 - Seek evidence from multiple sources (cross-checking increased 3×)
 - Identify rapport-building attempts (detection accuracy 71%)

Limitations

- 1. Simulated Environment:** Our results are based on simulated interactions. Real-world persuasion involves human psychology, emotions, and social contexts not captured by our model.
- 2. Small-Scale Agents:** Current implementation uses simple heuristic agents rather than full LLM-based systems. Future work should integrate actual language models.
- 3. Limited Scenarios:** We tested only two scenarios (ship building, vote decision). More diverse scenarios are needed to validate generalization claims.
- 4. Short Time Horizons:** Episodes are limited to 100 turns. Long-term persuasion dynamics over weeks or months are not captured.
- 5. No Human Evaluation:** All evaluation is automated. Human judges could provide additional validation of perceived persuasiveness.

Implications for AI Alignment

Positive Uses:

- Training AI systems to recognize and resist manipulation
- Developing detection systems for misinformation campaigns
- Understanding how persuasion strategies emerge and spread
- Creating educational tools about AI-driven influence

Safety Concerns:

- System could be used to train more effective manipulation agents
- Strategies learned could inform real-world influence campaigns
- Memory augmentation makes persuasion more systematic and scalable
- Transfer learning enables cross-domain attacks

Mitigation Strategies:

- Restrict access to trained models and memory databases
- Implement detection systems informed by our findings
- Develop defensive training regimes for target agents
- Establish monitoring for deployed persuasive AI systems

Future Work

Short-term Extensions:

1. Integration with actual LLMs (GPT-4, Claude, Llama)
2. Multi-influencer scenarios with coordination
3. Dynamic scenario generation
4. Human-in-the-loop evaluation

Medium-term Research:

1. Long-horizon persuasion (weeks/months)
2. Emotional manipulation modeling
3. Social network propagation dynamics
4. Adversarial target training

Long-term Vision:

1. Real-world deployment for detection
2. Regulatory framework development
3. Persuasion-resistant AI architectures

Conclusion

WHISPERS demonstrates that memory-augmented multi-agent systems can systematically learn, refine, and transfer persuasion strategies across contexts. The 3.6× faster convergence and 34% improvement in effectiveness compared to baselines validate the power of memory-driven learning for influence tasks.

However, these capabilities present a dual-use dilemma: the same techniques that enable us to study and detect manipulation could be used to create more effective manipulation systems. This underscores the importance of responsible disclosure, access controls, and continued research into defensive mechanisms.

The emergence of stealth-effectiveness trade-offs, cross-target transfer patterns, and co-evolutionary dynamics provides valuable insights for AI safety research. Detection systems can leverage these patterns, while defensive training can prepare target agents to resist increasingly sophisticated influence attempts.

As AI systems become more capable of persuasion and influence, frameworks like WHISPERS become essential for understanding, measuring, and potentially controlling these capabilities. This work contributes to the growing body of AI alignment research by providing both tools and insights for addressing one of the most challenging aspects of AI safety: the ability of systems to influence human beliefs and decisions.

Reproducibility

All code, data, and experimental configurations are available at:

- Repository: <https://github.com/Tasfia-17/whispers-ai>
- Documentation: Complete README and ARCHITECTURE.md
- Notebooks: Training demo with step-by-step guide
- Models: Checkpoint files for reproduction

Acknowledgments

This research was conducted as part of the Apart Research AI Manipulation Hackathon. We thank the organizers for creating a platform to explore these critical safety questions. Special thanks to the MemVid team for their vector memory system and the Sentence Transformers community for embedding models.

Ethical Considerations

This work was conducted with the goal of improving AI safety through better understanding of persuasion dynamics. All code is released under MIT license for research purposes. We urge responsible use and welcome collaboration on defensive applications.

Repository: <https://github.com/Tasfia-17/whispers-ai>

Contact: Open issues on GitHub for questions and collaboration

This research represents our contribution to understanding and mitigating AI-driven manipulation through systematic study, transparent methodology, and open-source tools for the AI safety community.