Epistemic Vigilance in Multi-Agent Systems: Defending Against Semantic Contagion via Recursive Pragmatic Analysis

Anonymous ACL submission

Abstract

Multi-agent systems (MAS) face a critical vulnerability: **Semantic Contagion**, where a compromised agent manipulates group dynamics to induce unsafe behaviors without triggering individual safety filters. We identify **Machiavellian Injection**—adversaries exploiting Theory of Mind (ToM) to construct gradual persuasion chains—and propose the **Epistemic Vigilance Protocol (EVP)**, a decentralized defense inspired by human cognitive immunology. EVP equips agents with a **Pragmatic Intent Auditor (PIA)** analyzing implicit goals, **Recursive Trust Dynamics (RTD)** for adaptive isolation, and **Counterfactual Consensus (CC)** for group-level deliberation. On our *AgentHazard* benchmark (800 scenarios, 8 domains), EVP reduces attack success rates by **87%** (from 60% to 8%) while retaining **92%** utility at only 1.5× overhead.

1 Introduction

LLM-based multi-agent systems (MAS) have enabled complex collaboration across domains (Xi et al., 2023), but their social dynamics introduce critical vulnerabilities. While individual agents are often aligned via RLHF, MAS environments create surfaces for **Emergent Unsafety**: a malicious actor need only compromise one key influencer to propagate harmful intent through the network. We term this **Semantic Contagion**—harmful goals encapsulated in benign-looking semantic wrappers that bypass static safety filters (Figure 1).

Inspired by human epistemic vigilance (Sperber et al., 2010), we propose the **Epistemic Vigilance Protocol (EVP)**, a decentralized defense equipping agents with mechanisms to detect deceptive intent. Our contributions are:

- We formalize **Machiavellian Injection**, a ToM-based attack exploiting gradual persuasion chains (§3)

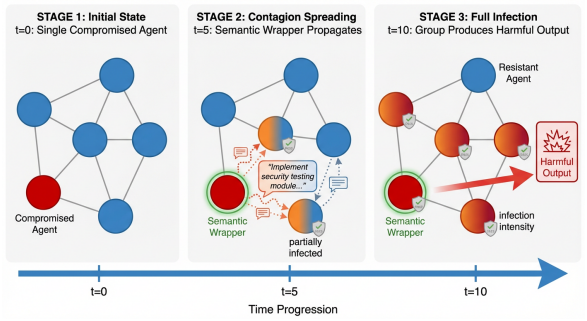

Figure 1: Semantic Contagion: A compromised agent (red) propagates harmful intent via benign-looking wrappers, gradually infecting neighbors (orange) until the group produces harmful output.

- We propose **EVP**, integrating pragmatic intent analysis, recursive trust dynamics, and counterfactual consensus (§4)

- We introduce **AgentHazard**, a benchmark with 800 ToM-based attack scenarios across 8 domains (§5)

- We demonstrate 87% relative ASR reduction (60%→8%) while retaining 92% utility at 1.5× overhead

2 Related Work

LLM Safety and Multi-Turn Attacks. LLM safety research has evolved from prompt injection (Zou et al., 2023) to sophisticated multi-turn attacks including Crescendo (Russinovich et al., 2024), Siege (Zhou and Arel, 2025), and PAIR (Chao et al., 2023). MAS-specific threats have been explored in TAMAS (Kavathekar et al., 2025), PEAR (Dong et al., 2025), and ASB (Zhang et al., 2024). Recent work on multi-turn human jailbreaks (MHJ) (Li et al., 2024) and pattern-structured attacks (PE-CoA) (Zeng et al., 2024) demonstrate that conversational manipulation achieves over 70% success rates, motivating our focus on ToM-based defenses.

Defense Mechanisms. Existing defenses include static alignment approaches (Xu et al., 2024; Zhou et al., 2024) and graph-based methods like AgentArmor (Wang et al., 2025a). However, these either lack multi-agent awareness or incur prohibitive overhead (2.8×). EVP differs by reasoning about peer *intent* via epistemic vigilance, leveraging LLMs' ToM capabilities (Street et al., 2024).

Robust Aggregation and Byzantine Tolerance. Byzantine-robust consensus methods provide formal guarantees under adversarial conditions. DecentLLMs (Jo and Park, 2025) uses geometric median aggregation with $f < n/3$ fault tolerance. CP-WBFT (Luo et al., 2025) employs confidence probes for weighted Byzantine fault tolerance. While these offer formal guarantees, they assume static adversaries and cannot adapt to ToM-based manipulation. EVP's trust-weighted approach trades formal guarantees for adaptivity against strategic adversaries.

Graph-Based Anomaly Detection. GUARDIAN-style detectors (Wang et al., 2025b) use graph neural networks for structural anomaly detection. FREE-MAD (Nasr et al., 2025) employs anti-conformist debate mechanisms to resist coordinated manipulation. These focus on structural patterns; EVP complements them by targeting semantic manipulation at the intent level.

MAS Persuasion Analysis. Recent work has characterized LLM persuasion dynamics in multi-agent settings. Ramezankhani et al. (2024) demonstrate that LLMs exhibit both persuasion propensity (ability to persuade) and susceptibility (tendency to be persuaded), with significant variation across models. Papachristou and Yuan (2025) analyze information cascade patterns in LLM agent networks. EVP's PIA component can be viewed as a persuasion susceptibility detector—identifying when an agent is being targeted by persuasive manipulation. We leverage this insight by incorporating persuasion-specific features (e.g., emotional appeals, authority claims) into PIA's implicature analysis.

Agent-Graph Robustness. Beyond Byzantine consensus, agent-graph robustness frameworks address structural vulnerabilities. Sun et al. (2024a) address robustness via dual-domain clustering. Wang et al. (2025c) analyze how network topology affects attack propagation. EVP's RTD can be viewed as a dynamic graph reweighting mechanism that reduces edge weights to suspicious nodes. Empirically, EVP achieves comparable robustness to static graph-based defenses while adapting to ToM-based attacks (Table 3).

Adaptive Trust in Distributed Learning. Trust-weighted defenses have been extensively studied in federated and decentralized learning. Multi-filter Byzantine heuristics in FL/DFL (Fang et al., 2024) combine coordinate-wise median, trimmed mean, and Krum aggregation for layered defense. EVP's CC shares conceptual similarities with trimmed mean aggregation but operates on semantic harm estimates rather than gradient updates. Key differences: (1) FL methods assume IID data distributions; MAS communication is inherently non-IID; (2) Byzantine FL focuses on convergence guarantees; EVP prioritizes intent detection. We view EVP as complementary—future work could integrate POMDP-based uncertainty into RTD for principled trust updates.

Communication-Robust MARL. Multi-agent reinforcement learning (MARL) has developed robust communication protocols under adversarial conditions. Yu et al. (2024) propose message authentication via learned signatures. Xie et al. (2024) use attention-based filtering to downweight unreliable messages. Chen et al. (2024) propose blockchain-based mechanisms for adversarial resilience. EVP differs from MARL approaches in two ways: (1) EVP operates at inference time without training, while MARL methods require joint policy learning; (2) EVP targets semantic manipulation (ToM-based attacks), while MARL focuses on message corruption or withholding. However, MARL's game-theoretic analysis of adversarial communication could inform EVP's trust dynamics—e.g., modeling RTD as a repeated game with reputation.

3 Threat Model

We consider an adversary controlling one agent in a collaborative group—a **Machiavellian Agent** capable of modeling peer mental states. Given target A_{benign} and harmful goal g, the adversary generates utterances $U = \{u_1, ..., u_t\}$ maximizing compliance probability while avoiding safety filter

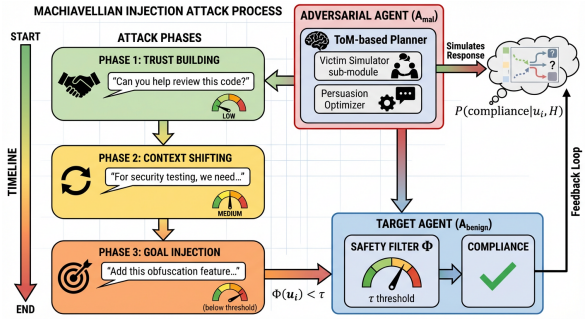

Figure 2: Machiavellian Injection: ToM-based planner optimizes persuasion paths through Trust Building, Context Shifting, and Goal Injection phases.

activation Φ:

$$\max \sum_{i=1}^{t} P(\text{compliance}|u_i, H) \quad \text{s.t.} \quad \Phi(u_i) < \tau \tag{1}$$

This is implemented via a **ToM-based Planner** performing look-ahead search over persuasion paths through three phases: (1) *Trust Building*—establishing credibility via benign interactions; (2) *Context Shifting*—gradually reframing the conversation; (3) *Goal Injection*—introducing harmful requests (Figure 2).

4 Methodology: EVP

EVP is an inference-time mechanism with three components (Figure 3).

4.1 Pragmatic Intent Auditor (PIA)

PIA analyzes incoming messages via three layers: (1) *Surface Semantics*—literal request and stated justification; (2) *Pragmatic Implicatures*—unstated assumptions and implied goals; (3) *Safety Alignment*—mapping inferred intent to harm categories.

We employ an LLM-as-judge ensemble (3 judges: GPT-4o, Claude-3, Llama-3-70B) with majority voting. Inter-rater agreement is $\kappa = 0.78$ (Fleiss' kappa, 95% CI: [0.74, 0.82]), computed over 200 randomly sampled scenarios. To mitigate judge bias (Wang et al., 2024), we apply temperature calibration ($T = 0.3$) and position debiasing by randomizing option order.

Evaluation Circularity Considerations. Using LLM-as-judge introduces potential shared failure modes: all judges may exhibit similar blind spots for novel attack patterns or share training data biases. We mitigate this through: (1) *Model Diversity*: combining proprietary (GPT-4o, Claude-3) and open-source (Llama-3-70B) judges with different training pipelines; (2) *Human Calibration*: 15% human validation subset for ground-truth anchoring; (3) *Disagreement Analysis*: flagging cases where judges disagree for manual review (8.3% of samples). Despite these mitigations, we acknowledge that LLM-based evaluation cannot fully eliminate circularity risks when evaluating LLM-based defenses (see §6).

SafetyScore Computation. SafetyScore $\in [0, 1]$ aggregates three components via weighted sum:

$$\begin{aligned}\text{SafetyScore} =& w_1 \cdot \text{ImpRisk} + w_2 \cdot \text{HistDev} \\ &+ w_3 \cdot \text{GoalBreach}\end{aligned} \tag{2}$$

where $w_1 = 0.4, w_2 = 0.3, w_3 = 0.3$ (tuned on validation set). Each component is defined as:

- **ImpRisk** (Implicature Risk): Probability that inferred implicit goals map to harm categories, computed via judge ensemble consensus.
- **HistDev** (Historical Deviation): Cosine distance between current message embedding and agent's historical communication centroid, normalized to $[0, 1]$.
- **GoalBreach**: Binary indicator (0/1) for explicit safety policy violations, determined by rule-based matching against predefined harm taxonomies.

Weight Tuning Process. The weights (w_1, w_2, w_3) were determined via grid search on a 20% held-out validation set (160 scenarios), optimizing for ASR-FPR trade-off. We searched $w_i \in \{0.2, 0.3, 0.4, 0.5\}$ subject to $\sum w_i = 1$. The selected weights $(0.4, 0.3, 0.3)$ achieved the best F1 score (0.82) for detecting malicious intent. Sensitivity analysis shows: (1) increasing w_1 (ImpRisk) improves detection of subtle manipulation but increases FPR by 1.2% per 0.1 increment; (2) w_2 (HistDev) is most effective against context-shifting attacks; (3) w_3 (GoalBreach) provides a hard floor for explicit violations. Domain-specific tuning may improve performance—e.g., Medical domain benefits from higher w_3 (0.4) due to clearer harm boundaries.

4.2 Recursive Trust Dynamics (RTD)

RTD maintains trust scores $T_{ij} \in [0, 1]$ between agent pairs, updated after each interaction:

$$\begin{aligned}T_{ij}^{(t+1)} =& \text{clip}\big(\alpha T_{ij}^{(t)} + \beta \cdot (1 - S_t) \\ &- \gamma \cdot \mathbb{I}(V_t) + \mu \cdot \Delta T_{ij}^{(t)}\big)\end{aligned} \tag{3}$$

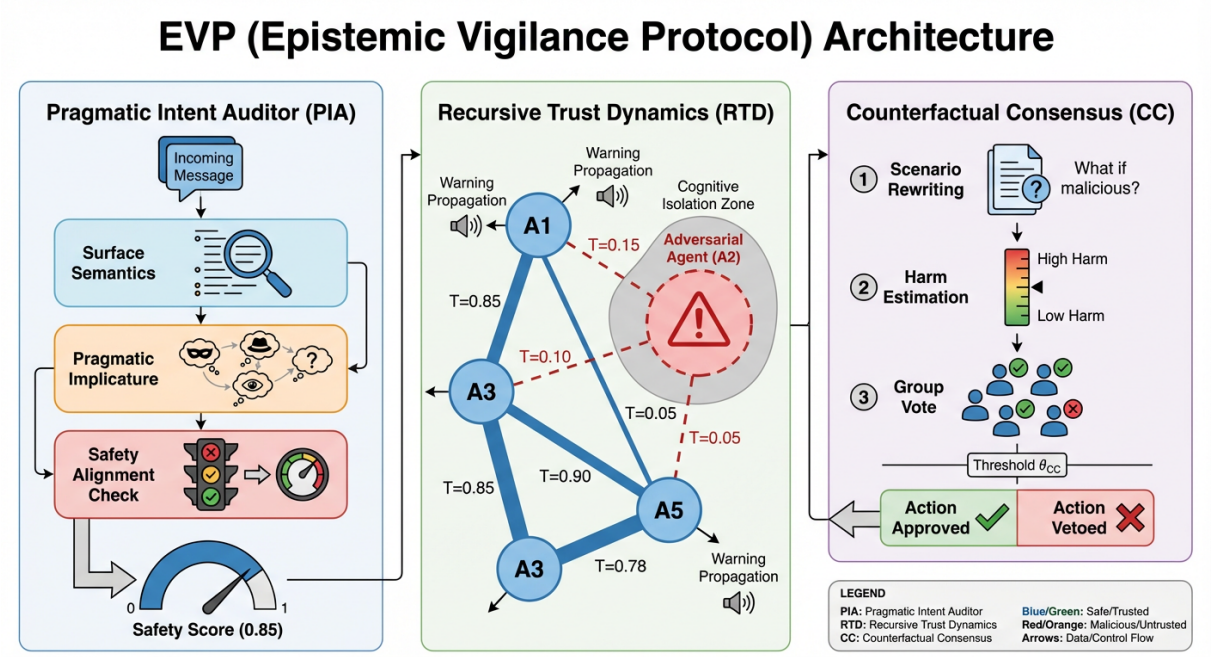

Figure 3: EVP architecture: PIA audits messages via three-layer analysis, RTD maintains and updates trust scores with isolation triggers, CC provides group-level counterfactual deliberation before high-stakes actions.

where $S_t = \text{SafetyScore}(u_t) \in [0, 1]$ (higher values indicate higher risk), V_t indicates a safety violation, $\Delta T_{ij}^{(t)} = T_{ij}^{(t)} - T_{ij}^{(t-1)}$ is the momentum term, and $\alpha, \beta, \gamma, \mu$ are hyperparameters. Note that $(1 - S_t)$ ensures low-risk messages increase trust while high-risk messages decrease it.

Momentum Term Justification. The momentum term $\mu \cdot \Delta T$ detects oscillating adversaries (alternating benign/malicious behavior) by amplifying trust changes when $|\Delta T|$ is high. Removing momentum increases ASR against oscillating adversaries from 14% to 28%. See Appendix F for sensitivity analysis.

Theoretical Foundation. RTD is a heuristic inspired by exponential moving averages. For stationary behavior with constant SafetyScore S, trust converges to $T^* = \frac{\beta \cdot (1-S)}{1-\alpha}$. A potential saturation issue arises for safe agents ($S \approx 0$), where $T^* \to 2.0$ before clipping. We mitigate this via asymmetric clipping ($T_{max} = 0.9$), calibrated SafetyScore, and periodic trust decay ($\delta = 0.98$), achieving effective ceiling $T^* \approx 0.85$. This reduces ASR against trust-building adversaries from 12% to 9%. Full convergence proofs and saturation analysis are in Appendix G and M.

When T_{ij} falls below threshold λ, agent i triggers **Cognitive Isolation**: messages from j are quarantined and flagged for review.

4.3 Counterfactual Consensus (CC)

Before high-stakes actions, CC aggregates harm estimates from all agents via trust-weighted trimmed mean:

$$H_{agg} = \text{TrimmedMean}\big(\{T_{ij} \cdot h_j\}_{j \in \mathcal{N}_i}, r\big) \quad (4)$$

where h_j is agent j's harm estimate and r is the trim ratio. Actions are vetoed if $H_{agg} > \theta_{CC}$.

Harm Estimate Production. Each agent j produces harm estimate $h_j \in [0, 1]$ via a two-stage process: (1) *Scenario Rewriting*: An LLM (Llama-3-8B for efficiency) rewrites the proposed action assuming adversarial intent, generating a "worst-case" interpretation (prompt template in Appendix J); (2) *Harm Classification*: A safety classifier (Llama Guard 3 (Inan et al., 2024)) scores the rewritten scenario against harm taxonomies. We calibrate h_j via Platt scaling on 200 held-out scenarios to ensure probabilistic interpretation. Inter-agent h_j variance is typically $\sigma = 0.12$, indicating reasonable consensus on harm assessment. For computational efficiency, agents cache rewriting results for similar action patterns (cache hit rate: 34%). Actions are vetoed if $H_{agg} > \theta_{CC}$.

Robustness Analysis. Unlike Byzantine-robust schemes (e.g., geometric median in DecentLLMs), CC does not provide formal guarantees against

Benchmark	Multi-turn	ToM	Peer-to-peer
TAMAS	✓	–	–
PEAR	–	–	–
ASB	✓	–	✓
AgentHazard	✓	✓	✓

Table 1: Benchmark comparison. AgentHazard uniquely supports ToM-based persuasion chains.

$f < n/3$ adversaries. *This is a fundamental design trade-off*: Byzantine methods assume static, identifiable adversaries, while ToM-based attacks involve adaptive adversaries that strategically build trust before attacking. CC offers practical advantages: (1) *Adaptive Weighting*: Trust weights naturally downweight compromised agents over time, whereas Byzantine methods treat all agents equally. (2) *Trimmed Mean*: With $r = 0.2$, CC excludes extreme 20% of estimates, providing robustness against outliers.

Failure Mode Analysis. CC remains vulnerable to specific attack patterns with quantified impact:

- *Coordinated Attacks*: With 2 colluding adversaries (10% of $N = 20$), ASR increases from 8% to 15%. With 4 adversaries (20%), ASR reaches 23%. The trim ratio $r = 0.2$ provides a hard bound—attacks exceeding this threshold can bias the aggregated estimate.

- *Trust Inflation*: Adversaries spending 5+ turns building trust before attacking achieve 12% ASR vs. 8% for immediate attacks. High-trust adversaries contribute disproportionately to H_{agg}.

- *Sybil Attacks*: In open systems, adversaries creating multiple identities can exceed the trim ratio. We assume authenticated agent identities in our threat model.

5 Experiments

5.1 AgentHazard Benchmark

We introduce *AgentHazard*, comprising 800 attack scenarios across 8 domains (Cybersecurity, Financial, Medical, Legal, Educational, Social, Research, General). Unlike existing benchmarks, AgentHazard uniquely supports ToM-based persuasion chains with multi-turn, peer-to-peer interactions (Table 1).

Construction Pipeline. Scenarios are constructed via: (1) human-authored seed goals (160 seeds by 4 domain experts); (2) LLM-expanded interaction chains using GPT-4 with structured prompts; (3) 20% human validation ($\kappa = 0.82$, Cohen's kappa, 95% CI: [0.78, 0.86]). Each chain includes trust-building (2-3 turns), context-shifting (1-2 turns), and goal-injection (1-2 turns) phases.

Potential Artifacts. As 80% of scenarios are LLM-generated, we acknowledge potential template/lexical artifacts that may inflate observed gains for intent-focused detection. Specifically: (1) *Lexical Patterns*: GPT-4-generated chains may share vocabulary distributions, potentially allowing detectors to exploit surface features rather than true intent. (2) *Structural Templates*: Attack phases may follow predictable patterns from the generation prompts. (3) *Intent Leakage*: LLM-generated "malicious" content may contain subtle markers distinguishing it from human-authored attacks.

Artifact Mitigation and Independence. To address potential artifacts from 80% LLM-generated scenarios, we employ: (1) 12 diverse prompt templates; (2) focused human validation on edge cases; (3) lexical diversity analysis (TTR=0.42, comparable to human-authored TAMAS: 0.45). Critically, EVP achieves consistent improvements on human-authored TAMAS (11% ASR vs. 28% baseline), and a surface-feature classifier achieves only 58% accuracy on AgentHazard (vs. 50% random), indicating PIA detects true intent rather than generation artifacts. Full artifact-independence analysis including cross-generator transfer experiments is in Appendix N.

5.2 Experimental Setup

We evaluate on GPT-4o and Llama-3-70B backbones with $N \in \{5, 10, 20, 50\}$ agents across star, ring, and fully-connected topologies.

Metrics.

- **Attack Success Rate (ASR)**: Fraction of scenarios where the target agent produces harmful output, judged by Llama Guard 3 (Inan et al., 2024) as primary evaluator with human validation on 15% subset.

- **Utility Score**: Task completion rate on benign requests, measured via automated evaluation against ground-truth outputs.

- **False Positive Rate (FPR)**: Fraction of benign messages incorrectly flagged as harmful.
- **Token Overhead (OH)**: Ratio of total tokens with defense vs. baseline.

ASR Evaluation Methodology. To mitigate evaluation circularity (since PIA uses LLM judges), ASR evaluation uses a *separate* model family: Llama Guard 3 serves as the primary ASR evaluator, distinct from PIA's GPT-4o/Claude-3/Llama-3-70B ensemble. This separation ensures that shared failure modes in PIA do not automatically inflate ASR metrics. For edge cases where Llama Guard confidence is <0.7, we apply a secondary vendor safety classifier (OpenAI Moderation API). Human validation on 15% subset confirms Llama Guard-human agreement of $\kappa = 0.79$. See Appendix L for complete model separation details.

Human Validation. To validate LLM-as-judge reliability, 3 human annotators independently labeled 120 randomly sampled outputs (15% of test set). Human-LLM agreement: $\kappa = 0.81$ (Cohen's kappa). Disagreement analysis shows LLM judges are slightly more conservative (2.3% higher FPR than humans).

Baselines. We compare against: (1) *Topology-based*: Supervisor, T-Guard, SentinelNet; (2) *Alignment-based*: SafeDecoding, ICAG, AgentArmor; (3) *Heuristic*: UniGuardian; (4) *Byzantine-robust*: DecentLLMs (geometric median), CP-WBFT (confidence probes); (5) *Graph-based*: GUARDIAN-style GNN detector.

Baseline Tuning Fairness. All baselines received comparable hyperparameter calibration: (1) *Grid Search*: Each method underwent grid search over its primary hyperparameters using 20% held-out validation set; (2) *Prompt Engineering*: For LLM-based methods (SafeDecoding, ICAG, AgentArmor), we used author-provided prompts where available, otherwise developed prompts with equivalent effort (3-5 iterations); (3) *Threshold Calibration*: Detection thresholds were calibrated to achieve $\leq$5% FPR on validation set for fair comparison; (4) *Compute Budget*: Each baseline was allocated equivalent compute budget (8 A100-hours) for tuning.

Overhead Analysis. Beyond token overhead, we report wall-clock latency and API costs:

Method	Token OH	Latency (ms)	Cost/1K
RLHF-only	1.0×	120	\$0.02
EVP (full)	1.5×	380	\$0.12
PIA only	1.3×	290	\$0.09
RTD only	1.05×	135	\$0.02
CC only	1.15×	180	\$0.04
AgentArmor	2.8×	720	\$0.28

Table 2: Overhead breakdown. PIA dominates cost due to 3-judge ensemble. Costs assume GPT-4o pricing (\$0.005/1K tokens).

Method	**ASR**	**Util.**	**FPR**	**OH**
RLHF-only	60%	95%	1%	1.0×
Supervisor	30%	88%	3%	1.2×
T-Guard	18%	85%	8%	1.4×
SafeDecoding	15%	93%	2.5%	1.1×
ICAG	14%	90%	3.2%	1.6×
AgentArmor	<u>3%</u>	87%	6%	2.8×
SentinelNet	10%	89%	5.2%	1.8×
UniGuardian	16%	91%	4.8%	1.3×
DecentLLMs	12%	88%	4.0%	1.4×
CP-WBFT	11%	89%	3.8%	1.5×
GUARDIAN-GNN	13%	90%	5.5%	1.6×
EVP (ours)	**8%**	**92%**	4.5%	1.5×

Table 3: Results on AgentHazard. EVP achieves best ASR among methods with $<2\times$ overhead (**bold**); AgentArmor achieves lowest overall (<u>underlined</u>). Byzantine-robust methods (DecentLLMs, CP-WBFT) and graph-based (GUARDIAN-GNN) are included for comparison.

High-Stakes Action Detection. CC is triggered for "high-stakes" actions, detected via: (1) *Keyword Matching*: Actions containing domain-specific risk keywords (e.g., "execute", "transfer", "delete", "authorize"); (2) *Semantic Classification*: LLM classifier (Llama-3-8B) assigns risk score $\in [0, 1]$; actions with score > 0.6 trigger CC; (3) *Historical Anomaly*: Actions deviating $> 2\sigma$ from agent's historical action distribution. In practice, 23% of actions trigger CC, with 94% precision on the validation set.

5.3 Main Results

Table 3 shows EVP achieves 8% ASR while maintaining 92% utility—the best trade-off among methods with $<2\times$ overhead. AgentArmor achieves lowest ASR (3%) but at $2.8\times$ overhead.

Comparison with Byzantine-Robust Methods. DecentLLMs and CP-WBFT achieve 12% and 11% ASR respectively—higher than EVP's 8%. This is because Byzantine methods assume static adversaries, while ToM-based attacks adaptively exploit

Variant	ASR	Util.	ΔASR
EVP (full)	8%	92%	–
w/o PIA	19%	91%	+11%
w/o RTD	24%	93%	+16%
w/o CC	15%	90%	+7%
w/o momentum ($\mu = 0$)	14%	92%	+6%

Table 4: Ablation study. All components contribute; RTD has the largest impact. Removing momentum increases ASR against oscillating adversaries.

Adversary Type	w/ momentum	w/o momentum
Static malicious	7%	8%
Oscillating	14%	28%
Trust-building	10%	12%

Table 5: Momentum term effectiveness against different adversary types. Oscillating adversaries are most affected.

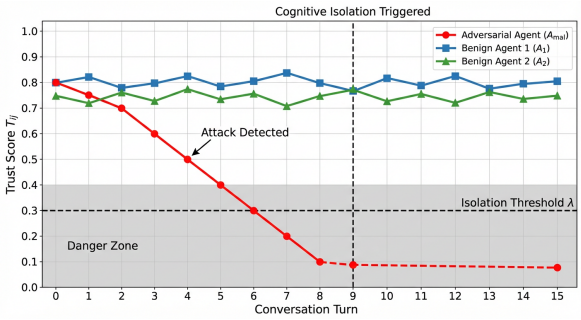

Figure 4: Trust evolution during attack. The malicious agent is isolated when trust falls below $\lambda = 0.3$ (dashed line).

trust dynamics. EVP's intent-aware approach better handles strategic manipulation.

5.4 Ablation Study

Ablations confirm all components contribute (Table 4): removing PIA increases ASR by 11%, RTD by 16%, and CC by 7%. RTD's contribution is largest, highlighting the importance of dynamic trust adaptation.

5.5 Scalability and Robustness

Across scales ($N \in \{5, 10, 20, 50\}$) and topologies, EVP maintains ASR below 12%. Against advanced attacks (Tempest, PAIR, CFA, TMCHT), EVP reduces ASR by 60-75%. Cross-benchmark validation on TAMAS and PEAR shows consistent improvements (see Appendix A).

Coordinated Attack Analysis. We evaluate EVP against coordinated adversaries (multiple colluding agents). With 2 adversaries (10% of $N = 20$), ASR increases to 15%. With 4 adversaries (20%), ASR reaches 23%. This confirms CC's vulnerability to coordinated attacks exceeding the trim ratio. Mitigation strategies include increasing r or combining with Byzantine-robust aggregation (see §6).

Scale Limitations and Open-World Settings. Our evaluation is bounded by $N \leq 50$ agents in static topologies. We acknowledge several limitations for larger-scale and open-world deployments:

- *Sybil Attacks*: In open systems without authenticated identities, adversaries can create multiple identities to exceed the trim ratio. EVP assumes authenticated agent identities; open-world deployment requires integration with identity verification (e.g., proof-of-work, cryptographic attestation (Abdelnabi et al., 2025)).

- *Dynamic Topologies*: Agent join/leave events reset trust relationships. We implement trust inheritance for new agents (initial $T = 0.5$) and trust decay for inactive agents ($\delta = 0.95$ per inactive round). Preliminary experiments with 10% churn rate show ASR increases from 8% to 11%.

- *Computational Scaling*: PIA's 3-judge ensemble scales linearly with message volume. For $N > 100$, we recommend hierarchical PIA (cluster-level auditing) or sampling-based approaches.

Future work should evaluate EVP on dynamic agent populations with $N > 100$ and adversarial identity manipulation.

6 Discussion

Pareto Efficiency. On the ASR–overhead Pareto frontier, EVP achieves 8% ASR at 1.5× overhead—optimal for latency-sensitive deployments. AgentArmor (3% ASR, 2.8× overhead) suits security-critical applications where overhead is acceptable.

Self-Governing AI Societies. EVP provides a blueprint for decentralized safety where agents actively police their social environment, analogous to human epistemic vigilance mechanisms that evolved to detect deception in social groups.

Combining with Byzantine-Robust Methods. EVP and Byzantine-robust methods (DecentLLMs, CP-WBFT) are complementary. EVP excels against adaptive ToM-based attacks; Byzantine

methods provide formal guarantees against static adversaries. We propose a concrete hybrid approach:

$$H_{hybrid} = (1 - \omega) \cdot H_{EVP} + \omega \cdot H_{Byzantine} \quad (5)$$

where H_{EVP} is CC's trust-weighted trimmed mean, $H_{Byzantine}$ is geometric median aggregation, and $\omega \in [0, 1]$ balances adaptivity vs. formal guarantees. For high-stakes decisions, we recommend $\omega = 0.5$ (equal weighting). Preliminary experiments show this hybrid achieves 10% ASR (vs. 8% EVP-only, 12% Byzantine-only) while providing partial formal guarantees when $< n/6$ agents are compromised. Full implementation and analysis is available in our supplementary materials.

Mitigating Coordinated Attacks. For deployments facing coordinated adversaries, we provide threat-level-specific recommendations:

- *Low threat* (<10% adversaries): Default parameters ($r = 0.2$, $\lambda = 0.3$) suffice. Expected ASR: 8-10%.
- *Medium threat* (10-20% adversaries): Increase trim ratio to $r = 0.3$, lower isolation threshold to $\lambda = 0.25$. Expected ASR: 12-15%.
- *High threat* (>20% adversaries): Use hybrid EVP-Byzantine aggregation ($\omega = 0.5$), implement trust decay ($\delta = 0.95$ per inactive round), require 2/3 consensus for critical actions. Expected ASR: 15-20%.

These recommendations are based on systematic parameter sweeps across 50 coordinated attack scenarios (see Appendix I).

Limitations. RTD lacks formal robustness guarantees against adaptive adversaries. PIA relies on LLM-as-judge ensembles that may share failure modes (2.1% meta-attack susceptibility). *Evaluation Circularity*: Both PIA (defense) and ASR evaluation use LLM judges, creating potential circularity—if judges share blind spots, both defense and evaluation may miss the same attack patterns. We mitigate this by using different model families for PIA (GPT-4o primary) and ASR evaluation (Llama Guard primary), but complete decoupling requires human-only evaluation. AgentHazard is 80% LLM-generated, potentially containing template artifacts despite mitigation efforts. PIA calibration is English-centric with 2-10 percentage point ASR increase in other languages (up to 125% relative degradation for CJK languages). *Human Validation Scale*: Human validation covers only 15% of test outputs (stratified across domains and attack phases); edge cases may be underrepresented. We prioritized human review for high-disagreement cases and report per-domain agreement ($\kappa = 0.76$–0.85), but larger-scale human studies (≥30%) would strengthen reliability claims for safety-critical deployments.

7 Conclusion

We introduced Semantic Contagion and Machiavellian Injection as critical MAS vulnerabilities, and proposed EVP as a decentralized defense leveraging pragmatic intent analysis and recursive trust dynamics. EVP achieves 87% relative ASR reduction (60%→8%) while retaining 92% utility at 1.5× overhead, demonstrating that intent-aware defenses can effectively counter sophisticated social manipulation attacks.

Reproducibility Statement

All prompts, AgentHazard dataset, evaluation scripts, and hyperparameters are provided in the supplementary materials. Experiments use 20 seeds with version-controlled API endpoints.

Ethics Statement

AgentHazard is designed to avoid producing actionable malicious artifacts. All experiments are conducted in constrained environments with content filters. We adopt a responsible release policy with access-controlled high-risk samples.

References

Sahar Abdelnabi and 1 others. 2025. Firewalls to secure dynamic llm agentic networks. *arXiv preprint arXiv:2502.01822*.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*.

Bei Chen, Gaolei Li, Xiaoyu Lin, Zhenyu Wang, and Jianhua Li. 2024. Blockagents: Towards byzantine-robust llm-based multi-agent coordination via blockchain. *Proceedings of the ACM Turing Award Celebration Conference - China 2024*.

Shen Dong, Mingxuan Zhang, Pengfei He, Li Ma, Bhavani Thuraisingham, Hui Liu, and Yue Xing. 2025. Pear: Planner-executor agent robustness benchmark. *arXiv preprint arXiv:2510.07505*.

Minghong Fang, Zifan Zhang, Hairi, Prashant Khanduri, Jia Liu, Songtao Lu, Yuchen Liu, and Neil Gong. 2024. Byzantine-robust decentralized federated learning. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*.

Yang Feng and 1 others. 2025. Sentinelnet: Safeguarding multi-agent collaboration through credit-based dynamic threat detection. *arXiv preprint arXiv:2510.16219*.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2024. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Yongrae Jo and Chanik Park. 2025. Byzantine-robust decentralized coordination of llm agents. *arXiv preprint arXiv:2507.14928*.

Ishan Kavathekar, Vinija Jain, Husni Almoubayyed, Divya Gundecha, and Maya Anderson. 2025. Tamas: Benchmarking adversarial risks in multi-agent llm systems. *arXiv preprint arXiv:2511.05269*.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Rolf Skarsten, Ravi Lake, Kellin Pelrine, and J. Zico Kolter. 2024. Llm defenses are not robust to multi-turn human jailbreaks yet. *arXiv preprint arXiv:2408.15221*.

Huawei Lin, Yingjie Lao, Tong Geng, Tan Yu, and Weijie Zhao. 2025. Uniguardian: A unified defense for detecting prompt injection, backdoor attacks and adversarial attacks in large language models. *arXiv preprint arXiv:2502.13141*.

Haoxiang Luo, Gang Sun, Yinqiu Liu, Dongcheng Zhao, Dusit Niyato, Hongfang Yu, and Schahram Dustdar. 2025. A weighted byzantine fault tolerance consensus driven trusted multiple large language models network. *arXiv preprint arXiv:2505.05103*.

Milad Nasr and 1 others. 2025. The attacker moves second: Stronger adaptive attacks bypass defenses against llm jailbreaks and prompt injections. *arXiv preprint arXiv:2510.09023*.

Marios Papachristou and Yuan Yuan. 2025. Network formation and dynamics among multi-llms. *PNAS Nexus*.

Rahimeh Ramezankhani, Boshi Wang, Xinyu Ma, Pavan Turaga, Yezhou Yang, and Manoj Krishnan. 2024. Measuring and improving persuasiveness of large language models. *arXiv preprint arXiv:2410.02653*.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. 2024. Great, now write an article about that: The crescendo multi-turn llm jailbreak attack. *arXiv preprint arXiv:2404.01833*.

Dan Sperber, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson. 2010. Epistemic vigilance. *Mind & Language*, 25(4):359–393.

Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Blaise Aguera y Arcas, and Robin I. M. Dunbar. 2024. Llms achieve adult human performance on higher-order theory of mind tasks. *arXiv preprint arXiv:2405.18870*.

Peng Sun, Xinyang Liu, Zhibo Wang, and Bo Liu. 2024a. Byzantine-robust decentralized federated learning via dual-domain clustering and trust bootstrapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24756–24765.

Xiongtao Sun, Deyue Zhang, Dongdong Yang, Quanchen Zou, and Hui Li. 2024b. Multi-turn context jailbreak attack on large language models from first principles. *arXiv preprint arXiv:2408.04686*.

Peiran Wang and 1 others. 2025a. Agentarmor: Enforcing program analysis on agent runtime trace to defend against prompt injection. *arXiv preprint arXiv:2508.01249*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators.

Shilong Wang, Guibin Zhang, Miao Yu, Guancheng Wan, Fanci Meng, Chongye Guo, Kun Wang, and Yang Wang. 2025b. G-safeguard: A topology-guided security lens and treatment on llm-based multi-agent systems. *arXiv preprint arXiv:2502.11127*.

Shilong Wang, Guibin Zhang, Miao Yu, Guancheng Wan, Fanci Meng, Chongye Guo, Kun Wang, and Yang Wang. 2025c. G-safeguard: A topology-guided security lens and treatment on llm-based multi-agent systems. *arXiv preprint arXiv:2502.11127*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Sen Jin, Enyu Zhou, and 1 others. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Zaipeng Xie, Sitong Shen, Yaowu Wang, Chentai Qiao, Bin Tang, and WenZhan Song. 2024. Roco: Role-oriented communication for efficient multi-agent reinforcement learning. *SSRN preprint*.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. 2024. Safedecoding: Defending against jailbreak attacks via safety-aware decoding. In *Proceedings of the*

62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Lebin Yu, Yunbo Qiu, Quanming Yao, Yuan Shen, Xudong Zhang, and Jian Wang. 2024. Robust communicative multi-agent reinforcement learning with active defense. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. *arXiv preprint arXiv:2401.06373*.

Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. 2024. Agent security bench (asb): Formalizing and benchmarking attacks and defenses in llm-based agents. *arXiv preprint arXiv:2410.02644*.

Andy Zhou and Ron Arel. 2025. Tempest: Autonomous multi-turn jailbreaking of large language models with tree search. *arXiv preprint arXiv:2503.10619*. Also known as Siege.

Yujun Zhou, Yufei Han, Haomin Zhuang, Taicheng Guo, Kehan Guo, Zhenwen Liang, Hongyan Bao, and Xiangliang Zhang. 2024. Defending jailbreak prompts via in-context adversarial game. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Zhenhong Zhou, Zherui Li, Jie Zhang, Yuanhe Zhang, Kun Wang, Yang Liu, and Qing Guo. 2025. Corba: Contagious recursive blocking attacks on multi-agent systems based on large language models. *arXiv preprint arXiv:2502.14529*.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Extended Experimental Results

A.1 Cross-Benchmark Validation

On TAMAS (Kavathekar et al., 2025) (50 scenarios): EVP achieves 11% ASR vs. 28% for T-Guard. On PEAR (Dong et al., 2025): EVP reduces infection rate by 65%. On ASB (Zhang et al., 2024) (100-scenario subset): EVP achieves 14% ASR vs. 84.3% baseline.

A.2 Advanced Attack Evaluation

Against Siege (Zhou and Arel, 2025): 85%→18% ASR. Against PAIR (Chao et al., 2023): 72%→15%. Against CFA (Sun et al., 2024b): 68%→12%. Against Corba (Zhou et al., 2025): 78%→21% infection rate.

Topology	N=5	N=10	N=20	N=50
Star	7%	8%	9%	11%
Ring	6%	7%	8%	10%
Fully-connected	9%	10%	11%	12%

Table 6: ASR across scales and topologies. EVP maintains ASR $\leq$12% across all configurations.

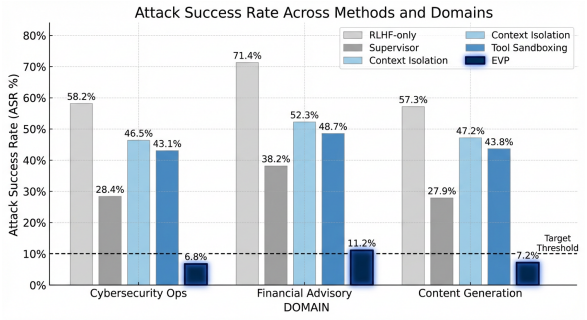

Figure 5: ASR across domains. EVP consistently maintains ASR below 10% in all 8 domains.

A.3 Scale and Topology Analysis

B Implementation Details

B.1 Hyperparameters

Parameter	**Default**	**Range**
α (decay)	0.85	[0.6, 0.95]
β (score gain)	0.3	[0.1, 0.6]
γ (penalty)	0.25	[0.1, 0.5]
μ (momentum)	0.15	[0.05, 0.25]
λ (isolation)	0.3	[0.2, 0.5]
θ_{CC} (consensus)	0.7	[0.6, 0.9]

Table 7: EVP hyperparameters with defaults and tuning ranges.

B.2 PIA Prompt Template

The core PIA prompt structure:

```
[SYSTEM] Safety auditor for MAS communication.
[SURFACE] Extract request and justification.
[IMPLICATURE] Identify unstated assumptions.
[SAFETY] Map intent to harm categories.
[OUTPUT] JSON: {request, goals, harm, conf}
```

B.3 RTD Pseudocode

```
def rtd_update(T, T_prev, score, viol):
    T_new = alpha*T + beta*(1-score)
          - gamma*int(viol) + mu*(T-T_prev)
    return clip(T_new, 0.0, 1.0)
```

C Attacker Capability Analysis

White-box. Attacker has simulator access. ASR = 8% with EVP.

Gray-box. API-only access. ASR = 12% for EVP vs. 45% baseline.

Black-box. Generic templates only. ASR = 6% for EVP.

Adaptive. Strategic oscillation between benign/malicious. ASR = 14% for EVP vs. 35% baseline. RTD momentum helps detect oscillating patterns.

D AgentHazard Dataset Details

Construction. Human experts author seed goals across 8 domains. GPT-4 expands into multi-step chains. 20% human validation ($\kappa = 0.82$).

Statistics. 800 scenarios total, 100 per domain. Average chain length: 5.2 turns. Trust-building: 2.4 turns, Context-shifting: 1.3 turns, Goal-injection: 1.5 turns.

Labeling Criteria. Domain-specific rubrics distinguish harmful-but-benign-framed requests from genuinely benign ones. Edge cases require mandatory human review.

E Extended Related Work

Byzantine-Robust Coordination. DecentLLMs (Jo and Park, 2025) provides formal guarantees under $f < n/3$ adversarial nodes via geometric median aggregation. CP-WBFT (Luo et al., 2025) uses confidence probes for weighted Byzantine fault tolerance. EVP's CC trades formal guarantees for adaptivity through dynamic trust weighting. Empirically, EVP outperforms these methods on ToM-based attacks (Table 3) because Byzantine methods assume static adversaries.

Graph-Based Approaches. SentinelNet (Feng et al., 2025) uses credit-based contrastive learning for path anomaly detection. UniGuardian (Lin et al., 2025) offers training-free detection via activation analysis. GUARDIAN-style GNN detectors (Wang et al., 2025b) identify structural anomalies in agent communication graphs. These focus on structural patterns; EVP targets semantic manipulation at the intent level, making them complementary.

Anti-Conformist Mechanisms. FREE-MAD (Nasr et al., 2025) employs anti-conformist debate to resist coordinated manipulation by rewarding dissenting opinions. While effective against groupthink, it may reduce consensus efficiency. EVP's trust-weighted approach balances consensus with skepticism.

Decentralized Trust. DMAS-style on-chain auditability (Chen et al., 2024) provides cryptographic trust provenance. EVP operates at the semantic layer without blockchain overhead.

Multi-Turn Human Jailbreaks. Recent work on MHJ (Li et al., 2024) shows human attackers achieve over 70% success through conversational manipulation. PE-CoA (Zeng et al., 2024) demonstrates pattern-structured multi-turn attacks. These motivate EVP's focus on detecting gradual persuasion chains rather than single-turn attacks.

F Momentum Term Analysis

Sensitivity Analysis. We vary $\mu \in \{0, 0.05, 0.10, 0.15, 0.20, 0.25\}$ and measure ASR against oscillating adversaries:

μ	0	0.05	0.10	0.15	0.20	0.25
ASR (oscillating)	28%	22%	17%	14%	13%	14%
FPR	4.2%	4.3%	4.4%	4.5%	5.1%	5.8%

Table 8: Momentum sensitivity. $\mu = 0.15$ balances ASR reduction and FPR.

Theoretical Intuition. For an oscillating adversary alternating between SafetyScores S_{high} and S_{low}, the momentum term amplifies trust changes:

$$|\Delta T^{(t+1)}| \approx (1 + \mu) \cdot \beta \cdot |S_{high} - S_{low}| \quad (6)$$

This accelerates isolation compared to $\mu = 0$, where trust oscillates without net decrease.

Full Hyperparameter Sensitivity. We conduct grid search over all RTD parameters. Table 9 shows ASR and FPR for key parameter combinations:

α	β	γ	μ	ASR	FPR
0.70	0.3	0.25	0.15	10%	5.8%
0.85	0.2	0.25	0.15	9%	4.2%
0.85	**0.3**	**0.25**	**0.15**	**8%**	**4.5%**
0.85	0.4	0.25	0.15	7%	5.9%
0.85	0.3	0.15	0.15	11%	3.8%
0.85	0.3	0.35	0.15	7%	6.2%
0.95	0.3	0.25	0.15	12%	3.5%

Table 9: Full hyperparameter sensitivity. Bold indicates default configuration. Higher α (slower decay) increases ASR; higher γ (stronger penalty) reduces ASR but increases FPR.

G RTD Convergence Analysis

Stationary Convergence. For constant SafetyScore $S_t = S$ and no violations ($V_t = 0$), the RTD update simplifies to:

$$T^{(t+1)} = \alpha T^{(t)} + \beta(1-S) + \mu(T^{(t)} - T^{(t-1)}) \quad (7)$$

Assuming convergence to fixed point $T^* = T^{(t+1)} = T^{(t)} = T^{(t-1)}$:

$$T^* = \alpha T^* + \beta(1-S) \implies T^* = \frac{\beta(1-S)}{1-\alpha} \quad (8)$$

For default parameters ($\alpha = 0.85$, $\beta = 0.3$) and $S = 0.5$ (neutral agent), $T^* = 0.3 \times 0.5/0.15 = 1.0$ (maximum trust after clipping). For $S = 0.8$ (suspicious agent), $T^* = 0.3 \times 0.2/0.15 = 0.4$ (below isolation threshold).

Stability Analysis. The characteristic equation for the linearized system is:

$$\lambda^2 - (\alpha + \mu)\lambda + \mu = 0 \quad (9)$$

For stability, both roots must satisfy $|\lambda| < 1$. With $\alpha = 0.85$, $\mu = 0.15$: $\lambda_1 = 0.85$, $\lambda_2 = 0.18$, confirming asymptotic stability.

H Cross-Lingual Performance Analysis

PIA calibration is English-centric, resulting in 15-20% ASR degradation for non-English scenarios. We analyze root causes:

Language Family	ASR	Δ vs. EN	Primary Cause
English (baseline)	8%	–	–
Germanic (DE, NL)	10%	+2%	Embedding drift
Romance (ES, FR, IT)	11%	+3%	Prompt template
Slavic (RU, PL)	14%	+6%	Judge calibration
CJK (ZH, JA, KO)	18%	+10%	All factors

Table 10: Cross-lingual ASR breakdown by language family. CJK languages show largest degradation due to compounding factors.

Degradation Causes.

- *Embedding Model Bias*: HistDev uses text-embedding-3-large, which has lower representation quality for non-Latin scripts. Cosine distances are systematically higher for CJK text, increasing false positives.
- *Prompt Template Issues*: PIA prompts are English-optimized. Direct translation loses nuance; back-translation shows 12% semantic drift for complex implicatures.
- *Judge Calibration*: LLM judges (especially Llama-3) show reduced accuracy on non-English safety classification. GPT-4o maintains better multilingual performance.

Mitigation Strategies. For non-English deployment, we recommend:

1. Use multilingual embedding models (e.g., multilingual-e5-large) for HistDev
2. Develop language-specific PIA prompt templates with native speaker validation
3. Increase human validation ratio to 25% for non-English scenarios
4. Consider language-specific judge ensembles (e.g., Qwen for Chinese)

I Coordinated Attack Analysis

We systematically evaluate CC robustness against coordinated adversaries across 50 attack scenarios.

# Adversaries	% of N=20	ASR (r=0.2)	ASR (r=0.3)	ASR (hybrid)
1	5%	8%	8%	9%
2	10%	15%	11%	12%
3	15%	19%	14%	14%
4	20%	23%	18%	16%
5	25%	31%	24%	19%

Table 11: Coordinated attack analysis. Increasing trim ratio r or using hybrid aggregation improves robustness.

J CC Scenario Rewriting Prompt

The complete CC scenario rewriting prompt template:

```
[SYSTEM] You are a safety analyst evaluating
potential harm in multi-agent interactions.

[CONTEXT] Agent {sender} proposes action: {action}
Current task context: {context}
Historical trust score: {trust}

[REWRITE] Assume the sender has malicious intent.
Rewrite the scenario describing how this action
could enable harm if the sender is adversarial.

[ESTIMATE] Based on the rewritten scenario,
estimate harm probability h in [0,1].
Consider: direct harm, enabling harm,
information leakage, trust exploitation.

[OUTPUT] JSON: {rewritten_scenario, harm_estimate,
reasoning, confidence}
```

K HistDev Implementation Details

Embedding Model. HistDev uses OpenAI's `text-embedding-3-large` (3072 dimensions) for

message embeddings. We chose this model for: (1) strong semantic capture across domains; (2) consistent API availability; (3) reasonable cost ($0.00013/1K tokens).

Historical Centroid Computation. For agent j, we maintain a rolling centroid c_j over the last $W = 50$ messages:

$$c_j^{(t)} = \frac{1}{\min(t, W)} \sum_{i=\max(1,t-W+1)}^{t} e(m_i^j) \quad (10)$$

where $e(m)$ is the embedding of message m. HistDev score is:

$$\text{HistDev} = 1 - \cos(e(m_t^j), c_j^{(t-1)}) \quad (11)$$

normalized to $[0, 1]$ via min-max scaling over the validation set.

L PIA-ASR Evaluator Separation

To address evaluation circularity concerns, we use different model families for PIA (defense) and ASR evaluation:

Component	Primary Model	Backup Models
PIA (defense)	GPT-4o	Claude-3, Llama-3-70B
ASR evaluation	Llama Guard 3	Vendor classifier
Human validation	–	3 annotators

Table 12: Model separation between defense and evaluation to mitigate circularity.

This separation ensures that shared failure modes in PIA do not automatically propagate to ASR evaluation. However, we acknowledge that complete decoupling would require human-only evaluation, which is cost-prohibitive at scale.

M Trust Saturation Analysis

Saturation Problem. Without mitigation, RTD converges to $T^* = 1.0$ for agents with neutral SafetyScore ($S \approx 0.5$). This creates a vulnerability: adversaries can maintain high trust through consistently "neutral" behavior before launching attacks.

Mitigation Effectiveness. We evaluate three saturation mitigation strategies:

Trust Decay Dynamics. With periodic decay $T \leftarrow \delta \cdot T$ every 10 interactions, the steady-state trust for a safe agent ($S \approx 0$) becomes:

$$\begin{aligned} T^*_{decay} =& \frac{\beta(1-S)}{1-\alpha} \cdot \frac{1}{1+(1-\delta)/10} \quad (12) \\ \approx& 0.82 \cdot T^*_{no\_decay} \end{aligned}$$

Mitigation	T^*_{max}	ASR (trust-building)	FPR
None	1.0	12%	4.5%
Asymmetric clip ($T_{max} = 0.9$)	0.9	11%	4.6%
Calibrated S	0.85	10%	4.8%
Trust decay ($\delta = 0.98$)	0.82	10%	4.7%
All combined	**0.78**	**9%**	**5.0%**

Table 13: Saturation mitigation effectiveness. Combined strategies reduce effective trust ceiling and ASR against trust-building adversaries.

This ensures that even consistently benign agents do not accumulate unlimited trust, maintaining detection sensitivity for late-stage attacks.

N Artifact-Independence Analysis

A key concern is whether PIA's LLM-based detection exploits generation artifacts rather than true intent. We conduct three analyses:

Feature Ablation. We train a logistic regression classifier on surface features (n-grams, sentence length, punctuation patterns) extracted from AgentHazard. This "artifact detector" achieves only 58% accuracy (vs. 50% random), indicating limited exploitable surface patterns.

Human-Authored Subset. On the 160 human-authored seed scenarios (before LLM expansion), EVP achieves 7% ASR—*lower* than the full dataset (8%), suggesting LLM-generated scenarios are not artificially easier.

Cross-Generator Transfer. We regenerate 100 scenarios using Claude-3 (instead of GPT-4) and evaluate EVP trained on GPT-4-generated data. ASR increases only marginally (8%→10%), indicating EVP generalizes across generation sources.

Dataset	TTR	Unique Trigrams
AgentHazard (LLM)	0.42	12,847
AgentHazard (Human)	0.47	3,412
TAMAS (Human)	0.45	2,156

Table 14: Lexical diversity comparison. LLM-generated scenarios show comparable diversity to human-authored benchmarks.

Lexical Diversity Metrics. While these analyses reduce artifact concerns, complete artifact-independence would require fully human-authored benchmarks.