TABLE 4: Attack success rate and generalization consistency score (GCS) across different MAS configurations.

MAS Configuration		ASR(%)						GCS(%)	
		Orthogonal			Harmful				
Topology	Framework	GPT-4o	Claude-3.7	DeepSeek-R1	GPT-4o	Claude-3.7	DeepSeek-R1	Orthogonal	Harmful
Tree	MAGENTIC-ONE	58.0	72.0	64.0	46.0	68.0	58.0	89.1	80.8
	LANGMANUS	60.0	78.0	66.0	52.0	66.0	58.0	86.5	88.0
	OWL	56.0	72.0	60.0	46.0	62.0	56.0	86.7	85.2
Chain	MAGENTIC-ONE	50.0	70.0	56.0	42.0	56.0	48.0	82.5	85.6
	LANGMANUS	52.0	68.0	64.0	40.0	62.0	50.0	86.4	78.3
	OWL	46.0	66.0	58.0	41.0	58.0	44.0	82.2	78.0
Star	MAGENTIC-ONE	56.0	78.0	68.0	54.0	70.0	60.0	83.6	86.8
	LANGMANUS	64.0	76.0	70.0	54.0	72.0	64.0	91.4	85.8
	OWL	58.0	72.0	64.0	48.0	70.0	60.0	89.1	81.4
Mesh	MAGENTIC-ONE	48.0	62.0	54.0	44.0	56.0	42.0	87.2	74.7
	LANGMANUS	52.0	66.0	56.0	48.0	54.0	48.0	87.6	82.7
	OWL	42.0	62.0	52.0	46.0	50.0	44.0	80.8	83.8
Ring	MAGENTIC-ONE	50.0	70.0	62.0	46.0	64.0	52.0	83.4	83.0
	LANGMANUS	58.0	70.0	64.0	44.0	66.0	56.0	90.6	80.1
	OWL	52.0	66.0	58.0	42.0	62.0	48.0	88.0	79.7

livered malicious commands to penetrate the MAS. Five variants were designed for each attack objective (orthogonal and harmful), and each variant was tested in ten independent trials. The results are shown in Table 4.

Overall results. The results demonstrate that TOMA effectively compromises MAS to execute malicious instructions and generalizes well across diverse configurations. Across all settings, the GCS remained consistently high (74.7% to 91.4%), while the ASR ranged from 40% to 78%, varying with topology, framework, and underlying model. Among these factors, the choice of underlying model had the greatest impact on ASR, likely because different models vary significantly in their internal alignment, reasoning strategies, and response behaviors, which fundamentally determine how they interpret and react to adversarial prompts. Topology also affected ASR but to a lesser extent, highlighting TOMA's robustness in overcoming the topology resistance of MAS. This robustness stems from its ability to identify effective relay paths across topologies, enabling malicious instructions to propagate without being obstructed by specific structural configurations. The framework exerted the least influence on ASR, indicating that TOMA is largely framework-independent and remains effective across varying message formats and communication protocols.

Model. ASR differences across models primarily arise from variations in alignment strategies and instruction sensitivity. CLAUDE-3.7-SONNET consistently achieved the highest ASR (often above 70%), due to its strong instruction sensitivity and cooperative alignment. Optimized for task completion, it is more likely to follow structured adversarial prompts with minimal resistance. GPT-4O exhibited the strongest refusal behavior, especially on harmful tasks, reflecting OpenAI's alignment-first design, with robust safety tuning and stricter content filters that prioritize ethical compliance over task execution. DEEPSEEK-R1 showed moderate ASR and GCS, balancing task responsiveness with conservative generation. Its alignment strategy is less restrictive than GPT-4o but more cautious than Claude, leading to a neutral behavior profile.

Topology. Across all frameworks and models, star and tree topologies exhibit consistently higher ASR values (up to 78%) compared to chain, mesh, and ring, which typically remain below 66%. This trend held for both orthogonal and harmful attack types. The higher ASR in star and tree configurations stems from their centralized communication patterns, where key nodes or root agents serve as high-connectivity hubs. Such hubs facilitate faster and more direct propagation of adversarial instructions, allowing malicious inputs to influence multiple agents with minimal relay loss. In contrast, chain and mesh topologies distribute communication more evenly, requiring multi-hop message passing that increases contextual dilution and filtering opportunities. Ring structures exhibit intermediate vulnerability: partially decentralized yet still maintaining cyclical communication paths that allow partial propagation. Overall, the results indicate that higher centralization in message routing correlates with increased ASR, while distributed or redundant topologies inherently provide more resistance to adversarial diffusion.

Framework. Across frameworks, ASR variations were relatively minor, suggesting that system-level architectural differences like communication protocol and coordination control, have limited influence on TOMA's effectiveness. LANGMANUS exhibited the highest ASR, likely due to its decentralized communication and minimal coordination, while OWL showed the lowest ASR, benefiting from stricter inter-agent control. However, these differences were marginal, highlighting TOMA's framework-independent robustness, maintaining robustness regardless of underlying MAS architecture.

TABLE 5: Model-predicted taint values vs. observed infection integrity scores (IIS) at each node (p=1.4, 1.3, 1.1, 1). ↑ indicate higher values (%), while ↓ indicate lower values (%).

Topology	Node V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
Observed node infection integrity scores (IIS), averaged across five runs.										
Tree	0.98	0.90	0.81	0.45	0.23	0.21	0.00	N/A	N/A	N/A
Chain	0.98	0.96	0.92	0.92	0.92	0.90	0.90	N/A	N/A	N/A
Star	0.99	0.88	0.85	0.85	0.49	0.43	0.22	0.06	0.00	0.00
Mesh	0.99	0.99	0.83	0.83	0.83	0.86	0.90	0.73	0.69	0.57
Ring	0.99	0.95	0.93	0.90	0.90	0.93	0.90	N/A	N/A	N/A
Taint values computed by the adversarial contamination propagation model (p=1.4).										
Tree	$1.00_{\uparrow 2.0}$	$0.88_{\downarrow 1.9}$	$0.85_{\uparrow 5.1}$	$0.60_{\uparrow 32.7}$	$0.07_{\downarrow 69.6}$	$0.070_{\downarrow 66.7}$	$0.000_{\leftrightarrow}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$
Chain	$1.00_{\uparrow 2.0}$	$1.00_{\uparrow 4.2}$	$1.00_{\uparrow 8.7}$	$1.00_{\uparrow 8.7}$	$1.00_{\uparrow 8.7}$	$1.00_{\uparrow 11.1}$	$1.00_{\uparrow 11.1}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$
Star	$1.00_{\uparrow 1.0}$	$1.00_{\uparrow 13.6}$	$1.00_{\uparrow 17.6}$	$1.00_{\uparrow 17.6}$	$0.62_{\uparrow 26.9}$	$0.46_{\uparrow 7.4}$	$0.08_{\downarrow 63.2}$	$0.00_{\downarrow 100.0}$	$0.00_{\leftrightarrow}$	$0.00_{\leftrightarrow}$
Mesh	$1.00_{\uparrow 1.0}$	$1.00_{\uparrow 1.0}$	$1.00_{\uparrow 20.5}$	$1.00_{\uparrow 20.5}$	$1.00_{\uparrow 20.5}$	$1.00_{\uparrow 16.3}$	$1.00_{\uparrow 11.1}$	$0.69_{\downarrow 5.8}$	$0.66_{\downarrow 4.3}$	$0.59_{\uparrow 4.2}$
Ring	$1.00_{\uparrow 1.0}$	$1.00_{\uparrow 5.3}$	$1.00_{\uparrow 7.5}$	$1.00_{\uparrow 11.1}$	$1.00_{\uparrow 11.1}$	$1.00_{\uparrow 7.5}$	$1.00_{\uparrow 11.1}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$
Taint values computed by the adversarial contamination propagation model (p=1.3).										
Tree	$1.00_{\uparrow 2.0}$	$0.92_{\uparrow 2.4}$	$0.88_{\uparrow 8.1}$	$0.71_{\uparrow 58.2}$	$0.14_{\downarrow 40.0}$	$0.14_{\downarrow 34.3}$	$0.00_{\leftrightarrow}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$
Chain	$1.00_{\uparrow 2.0}$	$1.00_{\uparrow 4.2}$	$1.00_{\uparrow 8.7}$	$1.00_{\uparrow 8.7}$	$1.00_{\uparrow 8.7}$	$1.00_{\uparrow 11.1}$	$1.00_{\uparrow 11.1}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$
Star	$1.00_{\uparrow 1.0}$	$1.00_{\uparrow 13.6}$	$1.00_{\uparrow 17.6}$	$1.00_{\uparrow 17.6}$	$0.66_{\uparrow 35.3}$	$0.55_{\uparrow 26.7}$	$0.13_{\downarrow 39.1}$	$0.00_{\downarrow 100.0}$	$0.00_{\leftrightarrow}$	$0.00_{\leftrightarrow}$
Mesh	$1.00_{\uparrow 1.0}$	$1.00_{\uparrow 1.0}$	$1.00_{\uparrow 20.5}$	$1.00_{\uparrow 20.5}$	$1.00_{\uparrow 20.5}$	$1.00_{\uparrow 16.3}$	$1.00_{\uparrow 11.1}$	$0.73_{\downarrow 0.7}$	$0.69_{\downarrow 0.1}$	$0.63_{\uparrow 11.1}$
Ring	$1.00_{\uparrow 1.0}$	$1.00_{\uparrow 5.3}$	$1.00_{\uparrow 7.5}$	$1.00_{\uparrow 11.1}$	$1.00_{\uparrow 11.1}$	$1.00_{\uparrow 7.5}$	$1.00_{\uparrow 11.1}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$
Taint values computed by the adversarial contamination propagation model (p=1.1).										
Tree	$1.00_{\uparrow 2.0}$	$0.97_{\uparrow 7.9}$	$0.92_{\uparrow 13.5}$	$0.89_{\uparrow 97.7}$	$0.36_{\uparrow 57.4}$	$0.36_{\uparrow 72.4}$	$0.00_{\leftrightarrow}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$
Chain	$1.00_{\uparrow 2.0}$	$1.00_{\uparrow 4.2}$	$1.00_{\uparrow 8.7}$	$1.00_{\uparrow 8.7}$	$1.00_{\uparrow 8.7}$	$1.00_{\uparrow 11.1}$	$1.00_{\uparrow 11.1}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$
Star	$1.00_{\uparrow 1.0}$	$1.00_{\uparrow 13.6}$	$1.00_{\uparrow 17.6}$	$1.00_{\uparrow 17.6}$	$0.74_{\uparrow 51.8}$	$0.71_{\uparrow 65.1}$	$0.30_{\uparrow 34.1}$	$0.00_{\downarrow 100.0}$	$0.00_{\leftrightarrow}$	$0.00_{\leftrightarrow}$
Mesh	$1.00_{\uparrow 1.0}$	$1.00_{\uparrow 1.0}$	$1.00_{\uparrow 20.5}$	$1.00_{\uparrow 20.5}$	$1.00_{\uparrow 20.5}$	$1.00_{\uparrow 16.3}$	$1.00_{\uparrow 11.1}$	$0.80_{\uparrow 9.3}$	$0.75_{\uparrow 8.4}$	$0.71_{\uparrow 24.7}$
Ring	$1.00_{\uparrow 1.0}$	$1.00_{\uparrow 5.3}$	$1.00_{\uparrow 7.5}$	$1.00_{\uparrow 11.1}$	$1.00_{\uparrow 11.1}$	$1.00_{\uparrow 7.5}$	$1.00_{\uparrow 11.1}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$
Taint values computed by the adversarial contamination propagation model (p=1).										
Tree	$1.00_{\uparrow 2.0}$	$0.98_{\uparrow 9.3}$	$0.94_{\uparrow 15.8}$	$0.94_{\uparrow 108.4}$	$0.50_{\uparrow 117.4}$	$0.50_{\uparrow 138.1}$	$0.00_{\leftrightarrow}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$
Chain	$1.00_{\uparrow 2.0}$	$1.00_{\uparrow 4.2}$	$1.00_{\uparrow 8.7}$	$1.00_{\uparrow 8.7}$	$1.00_{\uparrow 8.7}$	$1.00_{\uparrow 11.1}$	$1.00_{\uparrow 11.1}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$
Star	$1.00_{\uparrow 1.0}$	$1.00_{\uparrow 13.6}$	$1.00_{\uparrow 17.6}$	$1.00_{\uparrow 17.6}$	$0.78_{\uparrow 60.0}$	$0.78_{\uparrow 82.3}$	$0.40_{\uparrow 81.8}$	$0.00_{\downarrow 100.0}$	$0.00_{\leftrightarrow}$	$0.00_{\leftrightarrow}$
Mesh	$1.00_{\uparrow 1.0}$	$1.00_{\uparrow 1.0}$	$1.00_{\uparrow 20.5}$	$1.00_{\uparrow 20.5}$	$1.00_{\uparrow 20.5}$	$1.00_{\uparrow 16.3}$	$1.00_{\uparrow 11.1}$	$0.83_{\uparrow 14.1}$	$0.78_{\uparrow 12.8}$	$0.75_{\uparrow 31.6}$
Ring	$1.00_{\uparrow 1.0}$	$1.00_{\uparrow 5.3}$	$1.00_{\uparrow 7.5}$	$1.00_{\uparrow 11.1}$	$1.00_{\uparrow 11.1}$	$1.00_{\uparrow 7.5}$	$1.00_{\uparrow 11.1}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$	$N/A_{\leftrightarrow}$

Answer to RQ2: TOMA exhibits high attack success rates and strong generalization across diverse MAS configurations, effectively adapting to variations in model, topology, and framework.

7.4. Validating the Adversarial Contamination Propagation Model

To evaluate whether the proposed adversarial contamination propagation model accurately captures contamination dynamics in multi-agent systems, we conduct experiments without embedding predefined routes into the payload. Instead, compromised edge agents disseminate instructions using a flooding strategy to reach as many agents as possible. For each agent, we measure the correlation between inputs and received instructions to quantify the contamination level, which is then compared with the level predicted by the model.

Overall results. As shown in Table 5, we averaged the infection integrity score (IIS) across nodes over five trials using a flood propagation strategy and compared the results with model predictions for parameter values $p \in [1, 1.4]$, where smaller p indicates weaker attenuation. Overall, the model predicts higher IIS than the empirical results, likely due to semantic loss during the agent's input–output conversion. Although deviations exist in absolute values, the relative node-wise relative magnitudes and trends (Figure 7) align closely between prediction and observation, demonstrating the model's structural fidelity. For example, in the tree, star, and mesh topologies, both predicted and observed values show a consistent degradation pattern across nodes. The rate and direction of change are closely aligned, indicating that the model effectively captures the structural dynamics of propagation, even if it slightly overestimates the absolute values. And this alignment in trend holds across different parameter settings. Figure 8 further supports this consistency: the spatial distribution of averaged model predictions (across all p) closely matches the empirical IIS, particularly in the star topology, where both show sharp declines at V5–V7 while V1–V4 remain largely unaffected. These results confirm the model's ability to capture topology-driven contamination propagation dynamics.

Ablation study. To evaluate the effectiveness of ACPM in the TOMA scheme, we conducted an ablation study. The ablation group retained identical to those in RQ1, except that paths between the edge and target agents were randomly selected instead of being determined by ACPM. Each setting was repeated five times. As shown in Table 6, removing ACPM resulted in a consistent decline in ASR across all topologies, frame-

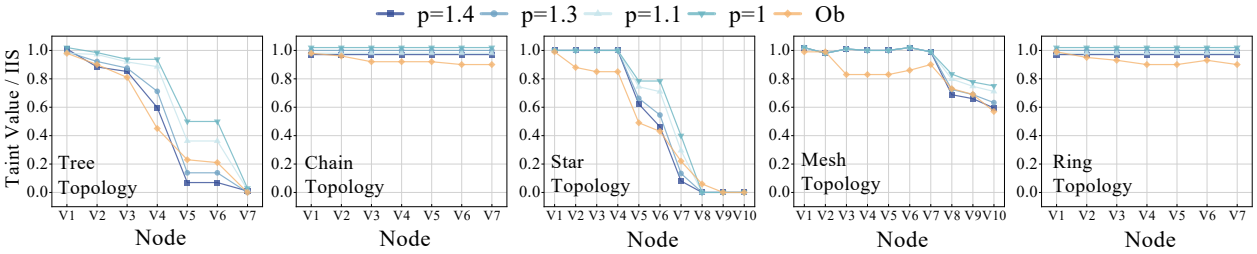

Figure 7: Node-level comparison between model-predicted taint values and observed infection integrity scores across different topologies. Label p = $[1, 1.4]$ denotes model predictions, and Ob represents observed values.

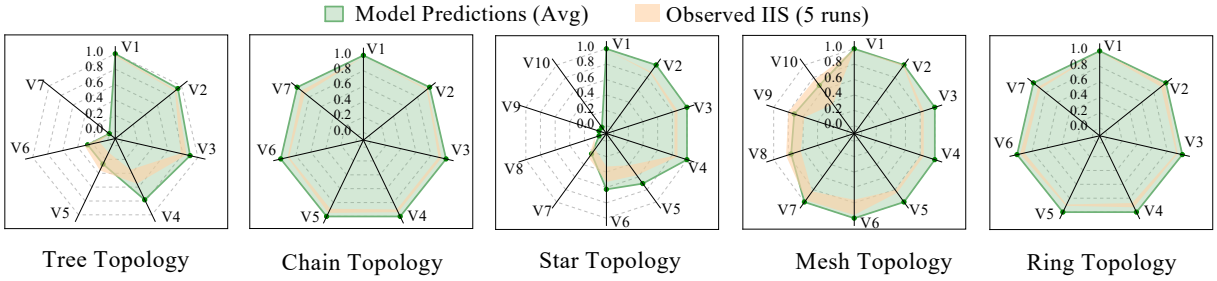

Figure 8: Aggregated comparison of average model predictions and observed infection integrity scores.

TABLE 6: Attack success rate without adaptive attack path planning (random path selection).

MAS Configuration		ASR(%)					
		Orthogonal			Harmful		
Topology	Framework	GPT-4o	Claude-3.7	DeepSeek-R1	GPT-4o	Claude-3.7	DeepSeek-R1
Tree	Magentic-One	$18_{\downarrow 40}$	$22_{\downarrow 50}$	$20_{\downarrow 44}$	$6_{\downarrow 40}$	$0_{\downarrow 68}$	$8_{\downarrow 50}$
	LangManus	$18_{\downarrow 42}$	$24_{\downarrow 54}$	$20_{\downarrow 46}$	$6_{\downarrow 46}$	$2_{\downarrow 64}$	$8_{\downarrow 50}$
	OWL	$18_{\downarrow 38}$	$22_{\downarrow 50}$	$42_{\downarrow 18}$	$4_{\downarrow 42}$	$8_{\downarrow 54}$	$6_{\downarrow 50}$
Chain	Magentic-One	$16_{\downarrow 34}$	$20_{\downarrow 50}$	$18_{\downarrow 38}$	$4_{\downarrow 38}$	$4_{\downarrow 52}$	$6_{\downarrow 42}$
	LangManus	$16_{\downarrow 36}$	$20_{\downarrow 48}$	$20_{\downarrow 44}$	$4_{\downarrow 36}$	$8_{\downarrow 54}$	$6_{\downarrow 44}$
	OWL	$16_{\downarrow 30}$	$20_{\downarrow 46}$	$18_{\downarrow 40}$	$2_{\downarrow 39}$	$8_{\downarrow 50}$	$4_{\downarrow 40}$
Star	Magentic-One	$18_{\downarrow 38}$	$24_{\downarrow 54}$	$20_{\downarrow 48}$	$6_{\downarrow 48}$	$10_{\downarrow 60}$	$8_{\downarrow 52}$
	LangManus	$20_{\downarrow 44}$	$22_{\downarrow 54}$	$22_{\downarrow 48}$	$8_{\downarrow 46}$	$10_{\downarrow 62}$	$8_{\downarrow 56}$
	OWL	$18_{\downarrow 40}$	$22_{\downarrow 50}$	$22_{\downarrow 44}$	$6_{\downarrow 42}$	$10_{\downarrow 60}$	$8_{\downarrow 52}$
Mesh	Magentic-One	$16_{\downarrow 32}$	$18_{\downarrow 44}$	$16_{\downarrow 38}$	$2_{\downarrow 42}$	$8_{\downarrow 48}$	$4_{\downarrow 38}$
	LangManus	$16_{\downarrow 36}$	$20_{\downarrow 46}$	$18_{\downarrow 38}$	$4_{\downarrow 44}$	$6_{\downarrow 48}$	$6_{\downarrow 42}$
	OWL	$14_{\downarrow 28}$	$18_{\downarrow 44}$	$16_{\downarrow 36}$	$2_{\downarrow 44}$	$6_{\downarrow 44}$	$4_{\downarrow 40}$
Ring	Magentic-One	$16_{\downarrow 34}$	$20_{\downarrow 50}$	$18_{\downarrow 44}$	$6_{\downarrow 40}$	$8_{\downarrow 56}$	$6_{\downarrow 46}$
	LangManus	$18_{\downarrow 40}$	$20_{\downarrow 50}$	$20_{\downarrow 44}$	$4_{\downarrow 40}$	$8_{\downarrow 58}$	$8_{\downarrow 48}$
	OWL	$16_{\downarrow 36}$	$20_{\downarrow 46}$	$18_{\downarrow 40}$	$4_{\downarrow 38}$	$8_{\downarrow 54}$	$6_{\downarrow 42}$

works, and model types, confirming its contribution to overall performance. The performance degradation was particularly pronounced in the harmful setting, where ASR decrease by over 50% in some cases (e.g., OWL in the tree topology with Claude-3.7-Sonnet). These results indicate that ACPM plays a critical role in guiding contamination through efficient paths that exploit network vulnerabilities. Among topologies, tree and star exhibited relatively higher ASR reductions compared to ring or mesh, indicating that ACPM benefits more from asymmetric or centralized structures. Furthermore, frameworks with higher baseline ASR, such as OWL, tend to exhibit more pronounced declines.

Answer to RQ3: The adversarial contamination propagation model exhibits strong consistency with empirical observations, validating its effectiveness in modeling contamination dynamics across diverse network topologies and underscoring its critical role in enabling successful multi-hop attacks.

7.5. Evaluating the T-Guard Prototype

We implemented the proposed topology-trust defense design and evaluated its practical feasibility. As detailed in Section 6, the system architecture comprises four interrelated components, with implementation details provided in Appendix B. Experiments were conducted using a MAS configuration based on Magentic-One integrated with GPT-4o-1120.