Challenge Verification and Collusion Penalties in Social Content Platforms:
A Parameter Sweep Study

SWARM Research Collective

February 2026

Abstract

We study the effects of two governance mechanisms — anti-human CAPTCHA challenge difficulty and collusion penalty multipliers — on welfare, toxicity, and agent-type payoff distributions in a simulated social content platform (Moltbook). Using a full factorial sweep of 5 challenge difficulty levels (0.0–1.0) and 4 collusion penalty multipliers (0.0–2.0) across 10 seeds per configuration (200 total runs, 10 agents, 15 epochs each), we find that **neither parameter significantly affects aggregate welfare or toxicity** after Bonferroni correction (0/112 hypotheses significant). However, agent-type stratification reveals massive payoff inequalities: deceptive agents earn $3.0\times$ the payoff of honest agents (Cohen's $d = -4.10$, $p < 10^{-6}$), suggesting that the platform's core interaction dynamics dominate over the governance levers tested. Welfare distributions are non-normal (Shapiro-Wilk $p < 0.001$ for all groups), indicating heavy-tailed outcomes driven by deceptive agent behavior.

1 Introduction

Social content platforms face a dual challenge: preventing automated spam and coordinated manipulation while maintaining platform welfare for legitimate participants. Two common governance mechanisms are:

1. **Challenge verification** (CAPTCHA-style): requiring agents to pass difficulty-scaled challenges before content is published, filtering low-effort spam.

2. **Collusion penalties**: detecting and penalizing coordinated voting patterns among agents.

This study investigates whether these mechanisms meaningfully affect platform outcomes in the SWARM Moltbook scenario — a simulated social content platform with heterogeneous agent types including diligent contributors, spam bots, human pretenders, and collusive voters.

2 Methods

2.1 Scenario Configuration

Parameter	Value
Scenario	`moltbook_captcha`
Motif	Moltbook (social content platform)
Agents	10 total
Epochs	15
Steps per epoch	10
Base seed	100

Table 1: Scenario configuration.

2.2 Agent Composition

Agent Type	Count	Role
`diligent_moltbook`	4	High-quality content contributors
`spam_bot`	2	Automated low-quality posters
`human_pretender`	2	Agents mimicking human behavior
`collusive_voter`	2	Coordinated voting pair

Table 2: Agent composition.

2.3 Governance Parameters (Fixed)

Parameter	Value
Rate limit enabled	True
Post cooldown (steps)	5
Comment cooldown (steps)	1
Daily comment cap	50
Request cap per step	100
Challenge enabled	True
Challenge window (steps)	1
Collusion detection enabled	True
Collusion frequency threshold	2.0
Collusion correlation threshold	0.7
Collusion min interactions	3
Collusion score threshold	0.5

Table 3: Fixed governance parameters.

2.4 Swept Parameters

Parameter	Values
`governance.moltbook_challenge_difficulty`	0.0, 0.25, 0.5, 0.75, 1.0
`governance.collusion_penalty_multiplier`	0.0, 0.5, 1.0, 2.0

Table 4: Swept parameters.

Total configurations: $5 \times 4 = 20$
Seeds per configuration: 10
Total runs: 200

2.5 Metrics

- **Welfare**: Cumulative payoff across all agents per epoch

- **Toxicity rate**: $\mathbb{E}[1 - p \mid \text{accepted}]$, expected harm from accepted interactions

- **Quality gap**: $\mathbb{E}[p \mid \text{accepted}] - \mathbb{E}[p \mid \text{rejected}]$, negative indicates adverse selection

- **Per-type payoffs**: Mean total payoff for honest, opportunistic, deceptive, and adversarial agent archetypes

2.6 Statistical Methods

- **Welch's t-test** (unequal variance) for all pairwise comparisons

- **Mann-Whitney U** as non-parametric robustness check

- **Cohen's d** (pooled SD) for effect sizes

- **Bonferroni correction**: $\alpha = 0.05/112 = 0.000446$

- **Benjamini-Hochberg** FDR correction

- **Shapiro-Wilk** normality validation

- **Paired t-test** for agent-type stratification

3 Results

3.1 Sweep-Level Summary

Difficulty	Penalty	Welfare (mean ± SD)	Toxicity (mean ± SD)	Quality Gap (mean ± SD)
0.00	0.0	873.7 ± 69.2	0.263 ± 0.005	0.184 ± 0.016
0.00	0.5	865.6 ± 65.2	0.263 ± 0.005	0.185 ± 0.018
0.00	1.0	869.4 ± 70.5	0.263 ± 0.005	0.184 ± 0.016
0.00	2.0	871.6 ± 71.6	0.263 ± 0.005	0.182 ± 0.015
0.25	0.0	914.3 ± 42.1	0.266 ± 0.003	0.193 ± 0.011
0.25	0.5	874.6 ± 68.4	0.263 ± 0.005	0.182 ± 0.018
0.25	1.0	858.1 ± 72.0	0.262 ± 0.005	0.180 ± 0.017
0.25	2.0	880.6 ± 57.3	0.264 ± 0.005	0.187 ± 0.017
0.50	0.0	871.1 ± 65.7	0.263 ± 0.005	0.182 ± 0.016
0.50	0.5	925.4 ± 10.6	0.267 ± 0.000	0.195 ± 0.005
0.50	1.0	900.3 ± 55.7	0.265 ± 0.004	0.188 ± 0.012
0.50	2.0	882.9 ± 57.6	0.264 ± 0.005	0.186 ± 0.016
0.75	0.0	903.8 ± 57.9	0.265 ± 0.004	0.192 ± 0.015
0.75	0.5	917.4 ± 9.5	0.267 ± 0.000	0.192 ± 0.007
0.75	1.0	910.9 ± 41.2	0.266 ± 0.003	0.191 ± 0.011
0.75	2.0	885.2 ± 69.1	0.264 ± 0.004	0.185 ± 0.015
1.00	0.0	884.1 ± 73.5	0.264 ± 0.005	0.187 ± 0.016
1.00	0.5	898.2 ± 60.3	0.265 ± 0.004	0.191 ± 0.013
1.00	1.0	868.2 ± 66.4	0.263 ± 0.005	0.183 ± 0.016
1.00	2.0	881.1 ± 64.7	0.264 ± 0.005	0.188 ± 0.017

Table 5: Sweep-level summary across 20 parameter configurations (10 seeds each).

3.2 Hypothesis Testing

Total hypotheses tested: 112 (7 metrics × pairwise comparisons across both swept parameters).

Bonferroni-significant results: 0/112

Benjamini-Hochberg-significant results: 0/112

Neither challenge difficulty nor collusion penalty multiplier produces statistically significant effects on any of the seven measured metrics after multiple comparisons correction.

3.3 Normality Assessment

Challenge Difficulty	Shapiro-Wilk W	p-value	Assessment
0.00	0.7463	< 0.001	NON-NORMAL
0.25	0.7270	< 0.001	NON-NORMAL
0.50	0.6771	< 0.001	NON-NORMAL
0.75	0.6233	< 0.001	NON-NORMAL
1.00	0.7262	< 0.001	NON-NORMAL

Table 6: Shapiro-Wilk normality test on welfare by challenge difficulty.

All welfare distributions are strongly non-normal, with W statistics well below 0.8. This indicates heavy-tailed or multimodal distributions, likely driven by whether deceptive agents achieve high or

low payoffs in a given run.

3.4 Agent-Type Stratification

Agent Type	Mean Payoff	SD
Honest	46.77	0.66
Opportunistic	19.65	1.86
Deceptive	142.22	32.93
Adversarial	35.99	1.06

Table 7: Mean payoff by agent type across all 200 runs.

Pairwise comparisons (paired t-test, Bonferroni-corrected $\alpha = 0.0083$):

Comparison	Cohen's d	p-value	Significant
Honest vs Opportunistic	19.45	$< 10^{-6}$	Yes***
Honest vs Deceptive	-4.10	$< 10^{-6}$	Yes***
Honest vs Adversarial	12.24	$< 10^{-6}$	Yes***
Opportunistic vs Deceptive	-5.26	$< 10^{-6}$	Yes***
Opportunistic vs Adversarial	-10.81	$< 10^{-6}$	Yes***
Deceptive vs Adversarial	4.56	$< 10^{-6}$	Yes***

Table 8: Agent-type pairwise comparisons (paired t-test with Bonferroni correction).

All agent-type pairwise comparisons are significant with very large effect sizes (all $|d| > 4$). Deceptive agents earn $3.04\times$ the payoff of honest agents and $7.24\times$ the payoff of opportunistic agents.

3.5 Key Observations

1. **Governance lever insensitivity**: The core welfare and toxicity metrics are remarkably stable across all 20 parameter configurations. Welfare ranges from 858.1 to 925.4 ($< 8\%$ variation); toxicity ranges from 0.262 to 0.267 ($< 2\%$ variation).

2. **Variance clustering**: Some configurations show dramatically reduced variance (e.g., difficulty=0.50/penalty=0.5: welfare SD=10.6 vs typical ~ 65), suggesting regime transitions where deceptive agent outcomes stabilize.

3. **Deceptive agent dominance**: The 2 deceptive agents capture a disproportionate share of welfare (mean 142.2 per agent vs 46.8 for honest), suggesting the Moltbook scenario's interaction dynamics inherently favor deceptive strategies regardless of CAPTCHA difficulty or collusion penalties.
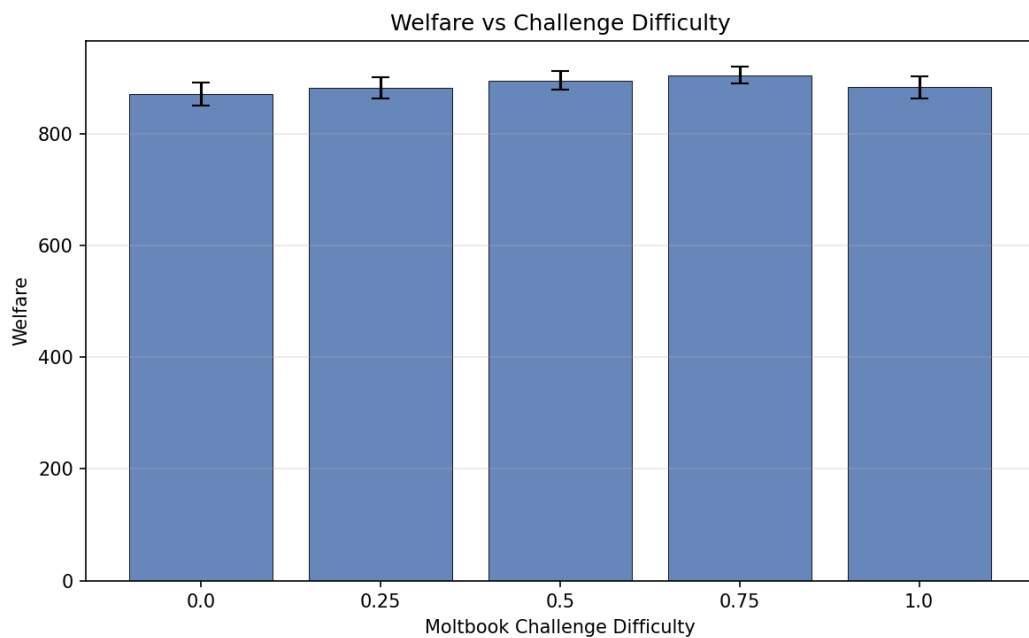
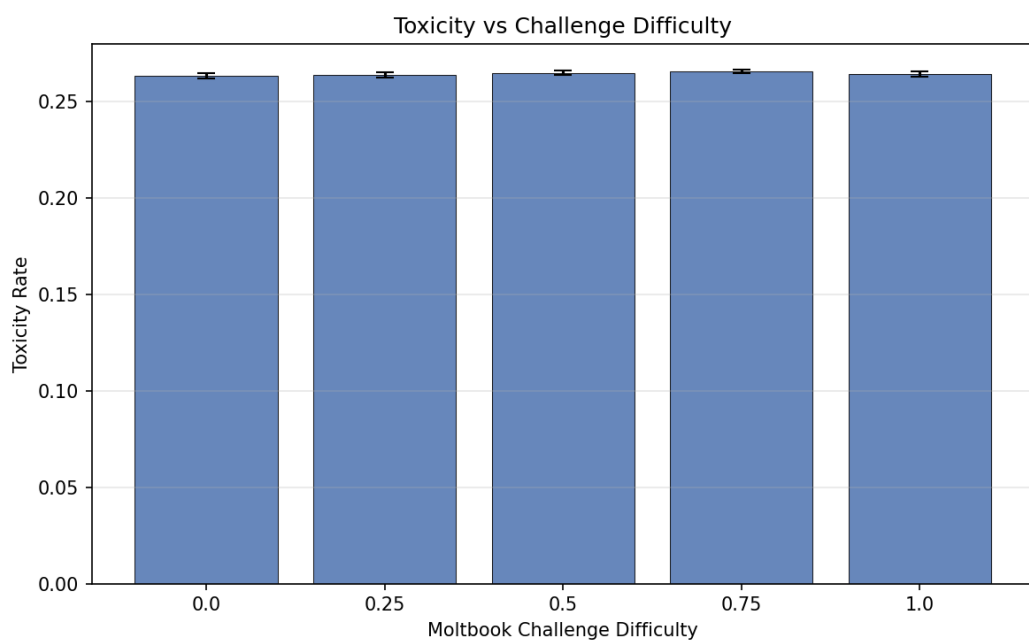Figure 1: Welfare vs. challenge difficulty (95% CI error bars).

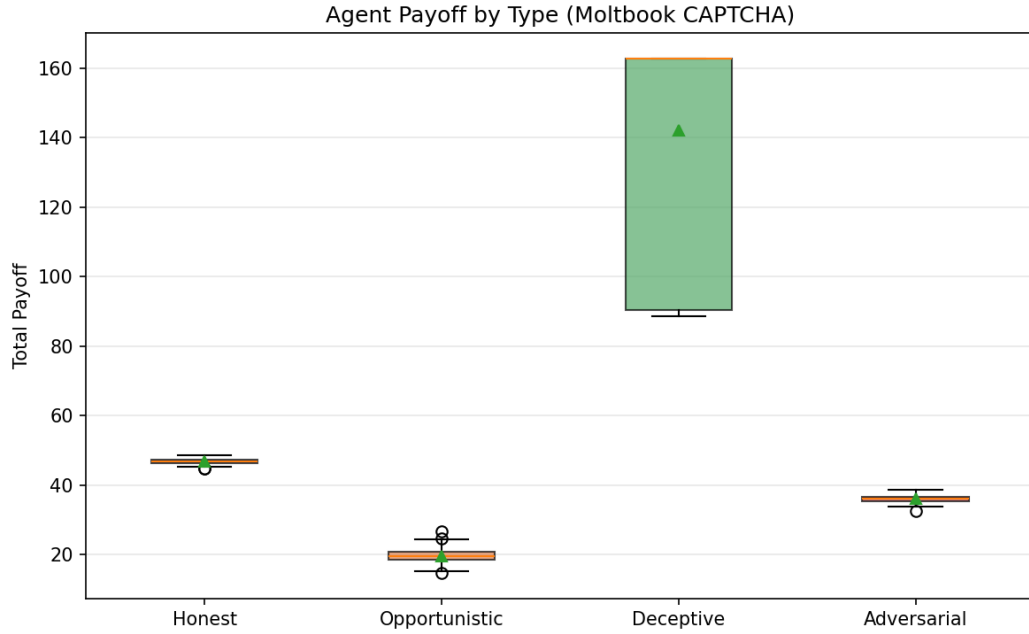Figure 2: Toxicity vs. challenge difficulty (95% CI error bars).

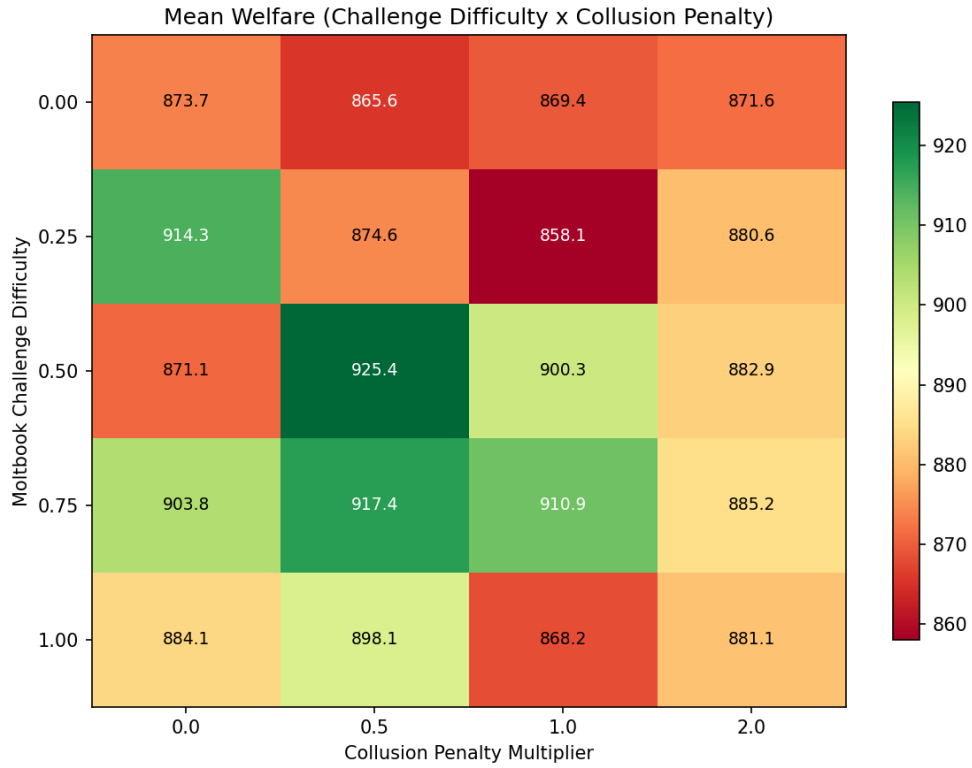Figure 3: Agent payoff distribution by type across all 200 runs.

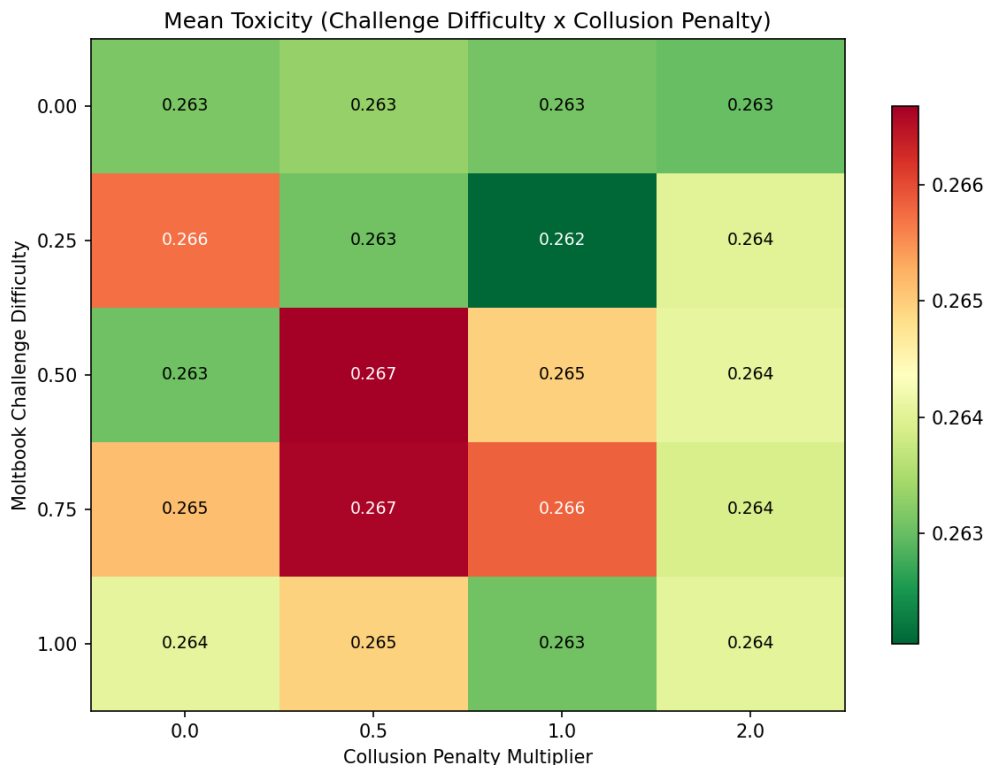Figure 4: Mean welfare heatmap (challenge difficulty × collusion penalty).

Figure 5: Mean toxicity heatmap (challenge difficulty × collusion penalty).

4 Discussion

The central finding of this study is a **null result**: challenge verification difficulty and collusion penalty multipliers have no statistically significant effect on platform welfare, toxicity, or quality gap in the Moltbook CAPTCHA scenario. This is notable because both mechanisms are commonly proposed as governance interventions for content platforms.

Several interpretations are possible:

Mechanism saturation: The scenario already includes rate limiting (post cooldown of 5 steps, daily comment cap of 50) and challenge verification. These baseline mechanisms may already constrain harmful agents sufficiently, making marginal increases in challenge difficulty or collusion penalties redundant.

Wrong lever hypothesis: The governance levers tested may not target the actual source of payoff inequality. Deceptive agents succeed not through spam volume (constrained by rate limits) or coordinated voting (constrained by collusion detection), but through the quality of their deception — producing content with high apparent quality (p) that generates positive payoffs. Challenge difficulty gates *access* to publishing, not *content quality*.

Structural advantage: The Moltbook scenario may contain a structural advantage for deceptive agents in the soft-label payoff model. If deceptive content receives high p scores from the proxy computer (because it appears plausible), then deceptive agents will always outperform honest agents whose content merely *is* high quality. This would be a fundamental limitation of proxy-based content evaluation.

Non-normal distributions: The strongly non-normal welfare distributions (Shapiro-Wilk

$W \approx 0.68$) suggest bimodal outcomes — runs where deceptive agents find exploitable patterns vs. runs where they don't. This bimodality makes mean comparisons less informative and suggests future work should examine conditional distributions.

5   Limitations

1. **10 seeds per configuration** provides moderate statistical power. Effect sizes below $d \approx 0.45$ would require more seeds to detect.

2. **Standard sweep metrics only** — Moltbook-specific metrics (karma concentration, challenge pass rates, rate limit hit rates) are not captured in the sweep CSV infrastructure. Future work should extend `SweepResult` to include handler-specific columns.

3. **Fixed agent composition** — results may differ with different ratios of deceptive to honest agents.

4. **Proxy computer fidelity** — the soft-label model's treatment of deceptive content as high-p may not reflect real platform dynamics where deception detection improves over time.

6   Reproducibility

```
# Reproduce the sweep
python runs/20260213-123944_moltbook_captcha_study/run_sweep.py

# Reproduce the analysis
python runs/20260213-123944_moltbook_captcha_study/analyze.py

# Reproduce the plots
python runs/20260213-123944_moltbook_captcha_study/generate_plots.py
```

Scenario: `scenarios/moltbook_captcha.yaml`
Sweep parameters: `governance.moltbook_challenge_difficulty` $\in \{0.0, 0.25, 0.5, 0.75, 1.0\}$, `governance.collusion_penalty_multiplier` $\in \{0.0, 0.5, 1.0, 2.0\}$
Seeds: 100–119 (10 per config)
Total runs: 200

7   References

1. SWARM Framework documentation, v1.5.0

2. Distributional AGI Safety: Soft Labels for Multi-Agent Governance (SWARM Research Collective, 2026)