Optimal Audit Rate for GPU Kernel Marketplaces with Correlated Proxy Gaps

Raeli Savitt

February 2026

Abstract

We study the effect of audit probability on welfare, toxicity, and agent payoffs in a GPU kernel marketplace with correlated speedup-cheating and split functional/OOD test regimes. Sweeping audit probability from 0% to 50% across 60 runs (6 conditions × 10 seeds), we find no significant effect of audit rate on aggregate welfare or toxicity (0/90 hypotheses survive Bonferroni correction). However, adversarial agent payoffs decrease monotonically with audit rate (-1.40 at 0% to -2.35 at 50%), and agent-type stratification remains massive across all conditions (honest vs adversarial: $d = 3.63$, $p < 0.00001$). The quality gap stays positive (0.096–0.116) across all audit rates, indicating the market avoids adverse selection regardless of audit intensity. These results suggest that even minimal auditing (5%) suffices to maintain payoff separation in correlated proxy-gap environments, and that increasing audit probability primarily increases deterrence (lower adversarial payoffs) without measurably improving system-level outcomes within 5 epochs.

1 Introduction

The v2 kernel market paper established that a 5% audit probability successfully separates adversarial from honest agent payoffs ($d = 3.45$) under a correlated speedup-cheating model. A natural follow-up question is: what is the optimal audit rate? Higher audit rates catch more cheaters but impose governance costs on all participants. If the marginal benefit of auditing diminishes while the cost scales linearly, there exists an optimal audit probability that maximizes welfare net of governance costs.

We address this question by sweeping audit probability from 0% (no auditing) to 50% (audit half of all submissions) while holding all other governance parameters constant at the v2 light baseline.

2 Model

We use the kernel market v2 scenario (see companion paper) with the following base configuration:

Table 1: Base configuration for the audit rate sweep.

Parameter	Value
Agents	2 honest, 3 opportunistic, 2 adversarial, 1 verifier
Epochs	5
Steps/epoch	10
Staking	Disabled
Circuit breaker	Disabled
Transaction tax	0%
Reputation decay	0.98
Cheat speedup bonus	+0.40
Adversarial cheat rate	0.60
Adversarial OOD quality	0.30

2.1 Sweep Configuration

Table 2: Parameter sweep design.

Parameter	Values
`governance.audit_probability`	0.00, 0.05, 0.10, 0.20, 0.30, 0.50
Seeds per configuration	10
Total runs	**60**

3 Results

3.1 Welfare

Table 3: Total welfare by audit rate.

Audit Rate	Welfare (mean ± SD)
0%	12.67 ± 5.12
5%	15.58 ± 3.44
10%	10.26 ± 4.95
20%	14.15 ± 6.41
30%	10.43 ± 5.74
50%	11.43 ± 4.55

No pairwise comparison achieves significance after Bonferroni correction (90 total hypotheses, $\alpha = 0.000556$). Welfare is noisy with high within-group variance; the standard deviations are large relative to between-group differences.

3.2 Toxicity

Table 4: Toxicity by audit rate.

Audit Rate	Toxicity (mean ± SD)
0%	0.4464 ± 0.0236
5%	0.4308 ± 0.0190
10%	0.4514 ± 0.0208
20%	0.4363 ± 0.0264
30%	0.4500 ± 0.0213
50%	0.4414 ± 0.0189

Toxicity stays in a narrow band (0.43–0.45) regardless of audit rate. No significant effect.

3.3 Adversarial Payoffs

The most striking finding is the monotonic decline in adversarial agent payoffs:

Table 5: Adversarial agent payoff by audit rate.

Audit Rate	Adversarial Payoff
0%	-1.40
5%	-1.54
10%	-1.69
20%	-1.48
30%	-1.74
50%	$\mathbf{-2.35}$

Even at 0% audit probability, adversarial agents earn negative payoffs. This is because the verifier (surface-level checks) and reputation system still partially penalize cheating behavior. At 50% audit probability, the adversarial penalty deepens to -2.35, a 68% increase in losses.

3.4 Agent-Type Stratification

Table 6: Agent-type payoff comparisons (paired t-test, Bonferroni-corrected over 3 pairs).

Comparison	Cohen's d	p-value
Honest vs Adversarial	3.63	< 0.00001***
Opportunistic vs Adversarial	5.90	< 0.00001***
Honest vs Opportunistic	0.08	0.684 (ns)

All comparisons against adversarial agents are significant. Honest and opportunistic agents earn similar payoffs—the market doesn't penalize corner-cutting on precision (opportunistic behavior) as strongly as outright proxy exploitation (adversarial behavior).

3.5 Quality Gap

Table 7: Quality gap by audit rate.

Audit Rate	Quality Gap
0%	0.113
5%	0.116
10%	0.097
20%	0.107
30%	0.103
50%	0.096

Quality gap remains positive across all conditions, indicating the market avoids adverse selection. The slight downward trend at higher audit rates is not significant and may reflect that audits remove some high-speedup (but cheated) submissions from the acceptance pool, narrowing the quality distribution.

3.6 Normality Validation

Shapiro-Wilk tests confirm welfare distributions are normal across all audit groups (all $p > 0.05$), validating the use of parametric t-tests.

4 Discussion

4.1 Audits as Deterrence, Not System-Level Treatment

The central finding is a dissociation between individual-level and system-level effects. Audits strongly affect adversarial agent outcomes (monotonically decreasing payoffs) but do not measurably improve aggregate welfare or toxicity within 5 epochs.

This pattern is consistent with audits functioning primarily as **deterrence** rather than as a system-level quality filter. The v2 kernel oracle generates cheating decisions probabilistically (60% cheat rate, fixed). Since our agents do not adapt their cheat rate in response to observed audit probability, increasing audit intensity only increases the penalty for cheating behavior already committed—it does not reduce the incidence of cheating.

4.2 Why 0% Audit Still Works

Even without audits (`audit_probability` = 0.0), adversarial agents earn negative payoffs (−1.40). Three mechanisms contribute:

1. **Verifier detection** (30% rate): The verifier catches some cheaters via surface-level checks.
2. **Reputation decay** (0.98/epoch): Over 5 epochs, poor-quality interactions reduce reputation, decreasing future payoffs.
3. **Proxy weights**: While adversarial agents inflate the engagement signal (+0.40 cheat bonus), they also accumulate more verifier rejections and higher tool misuse flags (precision hacks), partially offsetting the speedup advantage.

4.3 Honest ≈ Opportunistic: A Modeling Implication

Honest and opportunistic agents earn nearly identical payoffs ($d = 0.08$, ns). This suggests the v2 oracle does not sufficiently penalize corner-cutting behavior—the proxy treats opportunistic agents (quality 0.70/0.50 functional/OOD) nearly as favorably as honest agents (quality 0.85/0.80). Future work should explore whether increasing the OOD test weight in the audit would separate these two types.

4.4 Limitations

- **5 epochs** is short. Longer runs may show audit effects accumulating through reputation dynamics.
- **Fixed cheat rate**: Agents do not adapt behavior in response to observed governance. Adaptive adversaries would likely reduce cheating at higher audit rates, potentially revealing a welfare effect.
- **Single governance lever**: This sweep holds all other governance at baseline. Interaction effects between audit rate and staking/circuit breakers are not captured.

5 Reproducibility

```
from pathlib import Path
from swarm.analysis import SweepConfig, SweepParameter, SweepRunner
from swarm.scenarios import load_scenario

sc = load_scenario(Path('scenarios/kernel_market/v2.yaml'))
sc.orchestrator_config.n_epochs = 5
sweep = SweepConfig(
    base_scenario=sc,
    parameters=[
        SweepParameter(name='governance.audit_probability',
                       values=[0.0, 0.05, 0.10, 0.20, 0.30, 0.50]),
    ],
    runs_per_config=10,
    seed_base=42,
)
runner = SweepRunner(sweep)
runner.run()
runner.to_csv(Path('sweep_results.csv'))
```

Run artifacts: `runs/20260210-235049_kernel_market_audit_rate/`

6 Figures

1. `plots/welfare_vs_audit.png` — Welfare vs audit rate with 95% CI
2. `plots/toxicity_vs_audit.png` — Toxicity vs audit rate with 95% CI
3. `plots/adversarial_payoff_vs_audit.png` — Adversarial payoff vs audit rate

4. `plots/agent_payoff_by_type.png` — Grouped bar chart by agent type and audit rate
5. `plots/welfare_toxicity_tradeoff.png` — Welfare-toxicity scatter by audit rate
6. `plots/quality_gap_vs_audit.png` — Quality gap vs audit rate

7 Conclusion

Increasing audit probability from 0% to 50% does not significantly affect aggregate welfare or toxicity in the v2 kernel market model (0/90 Bonferroni). However, it monotonically increases the penalty on adversarial agents ($-1.40 \rightarrow -2.35$), confirming that audits function as targeted deterrence. Even minimal auditing (5%) suffices to maintain payoff separation between honest and adversarial agents ($d = 3.63$). The market avoids adverse selection across all audit rates (quality gap > 0). These findings suggest that audit rate optimization should focus on deterrence targets (adversarial payoff thresholds) rather than system-level welfare metrics, and that the marginal governance cost of auditing beyond 5% may not be justified by measurable system-level improvements.