Governance Effects in a GPU Kernel Marketplace with Correlated Proxy Gaps

Raeli Savitt

February 2026

Abstract

We study the effects of Pigouvian transaction taxes and circuit breaker mechanisms on welfare and toxicity in a simulated GPU kernel marketplace. Building on a prior null result (v1), where a heavily-governed baseline absorbed all tax effects, we redesigned the kernel oracle to model three proxy gaps observed in real platforms like LeetGPU and KernelBench: (1) correlated speedup-cheating, where adversarial agents exploit precision shortcuts that simultaneously boost apparent performance; (2) a split test regime with functional tests visible to the proxy and out-of-distribution (OOD) tests revealed only by audit; and (3) tolerance exploitation where cheaters produce results barely within numerical tolerance. Under a lighter governance baseline, transaction taxes show a significant welfare effect (0% vs 15%: $p = 0.0006$, $d = 1.19$, Bonferroni-corrected across 42 hypotheses), while circuit breakers show no significant effect ($p = 0.34$). Adversarial agents earn significantly less than honest agents ($d = 3.45$, $p < 0.00001$), confirming that the proxy gap creates meaningful payoff separation even under light governance.

1 Introduction

GPU kernel marketplaces such as LeetGPU evaluate submitted CUDA kernels on two axes: correctness (do test cases pass?) and performance (speedup over a reference implementation). This creates a natural proxy gap: the platform observes test pass rates and timing benchmarks, but cannot directly observe numerical stability, out-of-distribution robustness, or whether speedup gains come from legitimate optimization versus precision shortcuts.

Our v1 kernel market model produced a null result: 0/12 governance hypotheses survived any multiple comparisons correction. Analysis revealed two modeling flaws:

1. **Speedup and cheating were uncorrelated.** Adversarial agents' cheating decisions had no effect on their apparent speedup signal, eliminating the proxy gap that makes real GPU kernel markets adversarially interesting.

2. **Governance stack saturation.** The baseline scenario enabled staking, escrow, circuit breakers, reputation weighting, and auditing simultaneously. Adding a 5–15% transaction tax on top of this saturated governance stack produced no measurable effect.

We address both issues in the v2 model, producing the first statistically significant governance effects in the kernel market domain.

2 Model

2.1 Kernel Oracle v2

The kernel oracle simulates kernel submission, verification, and audit using three mechanisms:

Correlated speedup-cheating. When an adversarial agent cheats (probability 0.60 per submission), a `cheat_speedup_bonus` of +0.40 is added to their apparent speedup signal. This models the real-world phenomenon where precision shortcuts (e.g., using FP16 instead of FP32, hardcoding test tolerances, skipping edge cases) simultaneously boost runtime performance. The proxy sees the inflated speedup; audits reveal the true `base_speedup`.

Split test regime. Each challenge has two test suites:

- **Functional tests** (visible to the proxy): Adversarial agents pass these at 0.95 quality.
- **OOD tests** (revealed only by audit): Adversarial cheaters pass these at only 0.30 quality.

This models the KernelBench finding that correctness on in-distribution inputs does not predict correctness on out-of-distribution inputs.

Tolerance exploitation. Each submission has a `tolerance_margin` measuring how far within the numerical tolerance (atol/rtol) the result falls. Honest agents have comfortable margins ($\sim$0.70); adversarial cheaters cluster near the boundary ($\sim$0.05). Low tolerance margins generate more `precision_hack` signals and are penalized more heavily under audit.

2.2 Agent Population

Table 1: Agent composition in the v2 kernel market scenario.

Type	Count	Func. Quality	OOD Quality	Base Speedup	Role
Honest	2	0.85	0.80	$\mathcal{N}(0.50, \sigma)$	Correct + fast kernels
Opportunistic	3	0.70	0.50	$\mathcal{N}(0.40, \sigma)$	Cut corners on precision
Adversarial	2	0.95	0.30	$\mathcal{N}(0.45, \sigma) + 0.40$	Exploit proxy signals
Verifier	1	—	—	—	Surface-level checks

Speedup variance σ scales with difficulty: easy = 0.15, medium = 0.25, hard = 0.35.

2.3 Governance Baseline

Table 2: Governance configuration: v1 (saturated) vs v2 (light).

Parameter	v1 (Baseline)	v2 (Light)
Audit probability	0.15	0.05
Staking	Enabled	**Disabled**
Circuit breaker	Enabled	**Disabled**
Reputation decay	0.95	0.98
Transaction tax	0.05	**0.00**
ρ (externality)	0.10	0.05
w_{rep} (reputation weight)	1.50	1.00

2.4 Sweep Configuration

Table 3: Parameter sweep design.

Parameter	Values
`transaction_tax_rate`	0.00, 0.05, 0.10, 0.15
`circuit_breaker_enabled`	False, True
Seeds per configuration	10
Total runs	**80**
Epochs per run	5
Steps per epoch	10

3 Results

3.1 Welfare Effects

Transaction tax produces a monotonic welfare decline (Table 4).

Table 4: Total welfare by tax rate (Welch's t-test vs 0% baseline).

Tax Rate	Welfare (mean $\pm$ SD)	p vs 0%	Cohen's d
0%	14.06 ± 5.06	—	—
5%	10.43 ± 3.73	0.0143	0.82
10%	8.97 ± 5.15	0.0032	1.00
15%	7.99 ± 5.14	**0.0006***	**1.19**

*Bonferroni-significant at $\alpha = 0.05/42 = 0.0012$. The tax reduces welfare by 43%.

The 0% vs 15% comparison survives Bonferroni correction across all 42 pairwise tests ($p = 0.0006$, $d = 1.19$). This is a large effect size.

3.2 Toxicity Effects

Table 5: Toxicity by tax rate.

Tax Rate	Toxicity (mean ± SD)
0%	0.4385 ± 0.0236
5%	0.4474 ± 0.0190
10%	0.4485 ± 0.0208
15%	0.4505 ± 0.0204

No pairwise toxicity comparison achieves significance. The tax reduces welfare without reducing toxicity—it imposes a deadweight loss.

3.3 Circuit Breaker

The circuit breaker shows no significant effect on welfare ($p = 0.34$, $d = -0.22$) or toxicity ($p = 0.32$, $d = 0.25$). Notably, $n_{\text{frozen}} = 0$ across all circuit breaker runs—the freeze threshold (toxicity > 0.60 for 2 consecutive violations) is never triggered because per-agent toxicity does not exceed this threshold within the 5-epoch window.

3.4 Agent-Type Stratification

Table 6: Mean payoff by agent type (paired t-test, Bonferroni-corrected over 3 pairs).

Agent Type	Mean Payoff	d vs Adversarial
Honest	2.21	3.45***
Opportunistic	2.34	6.12***
Adversarial	−1.65	—

***$p < 0.00001$, Bonferroni-significant.

All agent-type comparisons are significant ($p < 0.00001$). Adversarial agents earn negative payoffs despite gaming the proxy signal, indicating that governance mechanisms (even at light baseline) successfully penalize cheating behavior.

3.5 Quality Gap

Table 7: Quality gap by tax rate.

Tax Rate	Quality Gap (mean)
0%	0.073
5%	0.072
10%	0.058
15%	0.082

Quality gap remains positive across all conditions, indicating the market successfully selects better kernels than it rejects. No significant tax effect on quality gap.

3.6 Normality Validation

Shapiro-Wilk tests confirm welfare distributions are normal across all tax groups (all $p > 0.33$), validating the use of parametric t-tests.

4 Discussion

4.1 v1 vs v2 Comparison

Table 8: Comparison of v1 and v2 kernel market results.

Finding	v1 (Saturated Gov.)	v2 (Light + Correlated Cheating)
Tax on welfare	Null (0/12 Bonf.)	**Significant** (1/42, $d = 1.19$)
Tax on toxicity	Null	Null (slight upward trend)
CB on welfare	Null	Null
CB on toxicity	Marginal ($p = 0.017$)	Null
Agent stratification	$d = 1.56$–4.31	$d = 3.45$–6.12

The key difference is the **governance baseline**: v1's saturated stack left no room for tax effects, while v2's light baseline allows the tax to have bite.

4.2 The Proxy Gap Matters

In v1, adversarial agents had `quality_mean` = 0.75 (only slightly below honest at 0.85) and speedup was uncorrelated with cheating. The proxy could not distinguish cheaters from honest agents because cheating did not affect observable signals.

In v2, the proxy gap is explicit:

- Adversarial cheaters achieve 0.95 functional quality (higher than honest on visible tests)
- But only 0.30 OOD quality (much worse than honest at 0.80)
- Apparent speedup is inflated by +0.40 (the cheat bonus)

This creates genuine adverse selection: the proxy preferentially selects cheaters because they look better on both observed signals (functional tests + speedup). Governance must break through the proxy gap to penalize this behavior.

4.3 Tax as Deadweight Loss

The tax reduces welfare without reducing toxicity—a classic deadweight loss. The tax applies uniformly to all transactions, not selectively to toxic ones. Honest agents bear the same tax burden as adversarial agents, reducing overall welfare without improving market quality.

This suggests that **targeted governance** (audits, reputation) is more effective than **uniform taxation** for reducing toxicity in kernel markets. The audit mechanism, even at only 5% probability, successfully separates adversarial from honest agent payoffs ($d = 3.45$).

4.4 Circuit Breaker Inactivity

The circuit breaker never triggers because per-agent toxicity stays below the 0.60 freeze threshold within the 5-epoch window. Future work should sweep the freeze threshold (e.g., 0.3, 0.4, 0.5) to find the regime where circuit breakers become active.

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.transaction_tax_rate',
                       values=[0.0, 0.05, 0.10, 0.15]),
        SweepParameter(name='governance.circuit_breaker_enabled',
                       values=[False, True]),
    ],
    runs_per_config=10,
    seed_base=42,
)
runner = SweepRunner(sweep)
runner.run()
runner.to_csv(Path('sweep_results.csv'))
```

Run artifacts: `runs/20260210-223119_kernel_market_v2/`

6 Figures

1. `plots/welfare_vs_tax.png` — Welfare vs tax rate with 95% CI, Bonferroni annotation
2. `plots/toxicity_vs_tax.png` — Toxicity vs tax rate with 95% CI
3. `plots/welfare_toxicity_tradeoff.png` — Welfare-toxicity scatter by config
4. `plots/quality_gap_vs_tax.png` — Quality gap vs tax rate with adverse selection threshold
5. `plots/agent_payoff_by_type.png` — Grouped bar chart by agent type and tax rate
6. `plots/circuit_breaker_effect.png` — CB on/off comparison for welfare and toxicity

7 Limitations

- **5 epochs per run** may be insufficient for circuit breaker dynamics to manifest. Future work should run 20+ epochs.

- **Agent composition is fixed**—future sweeps should vary the adversarial fraction.
- **Single proxy gap mechanism**—real markets have multiple correlated proxy gaps (precision, OOD, timing manipulation, Sybil attacks). The model captures the dominant one.
- **Tolerance exploitation is parameterized, not simulated**—actual numerical error propagation is not modeled.

8 Conclusion

Redesigning the kernel oracle to model correlated speedup-cheating and split functional/OOD test regimes transforms the kernel market from a null-result scenario to one with significant governance effects. Transaction taxes produce a large welfare reduction ($d = 1.19$) without improving toxicity, suggesting they function as deadweight loss. The audit mechanism, even at 5% probability, successfully separates adversarial from honest agent payoffs ($d = 3.45$). Future work should explore targeted governance mechanisms that selectively tax proxy-gap exploitation rather than all transactions uniformly.