Recursive Reasoning in Multi-Agent Systems: Strategic Depth as a Distributional Safety Risk

Raeli Savitt

February 2026

Abstract

We study the distributional safety implications of embedding strategically sophisticated agents—modeled as Recursive Language Models (RLMs) with level-k iterated best response—into multi-agent ecosystems governed by soft probabilistic labels. Across three pre-registered experiments ($N = 30$ seeds total, 26 statistical tests), we find three counter-intuitive results. **First**, deeper recursive reasoning *hurts* individual payoff (Pearson $r = -0.75$, $p < 0.001$, 10/10 tests survive Holm correction), rejecting the hypothesis that strategic depth enables implicit collusion. **Second**, memory budget asymmetry creates statistically significant but practically modest power imbalances (3.2% spread, $r = +0.67$, $p < 0.001$, 11/11 survive Holm). **Third**, fast-adapting RLM agents outperform honest baselines in small-world networks (Cohen's $d = 2.14$, $p = 0.0001$) but *not* by evading governance—rather by optimizing partner selection within legal bounds. Across all experiments, honest agents earn 2.3–2.8× more than any RLM tier in complete networks, suggesting that strategic sophistication is currently a net negative in SWARM-style ecosystems with soft governance. All p-values survive Holm–Bonferroni correction at the per-experiment level.

1 Introduction

The intersection of recursive reasoning and multi-agent safety is understudied. Most alignment research treats agents as either aligned or misaligned in a binary sense. The SWARM framework instead uses *soft probabilistic labels*—each interaction receives a probability p of being beneficial—enabling a richer distributional analysis of safety outcomes.

Recursive Language Models (RLMs) represent a class of agents that apply iterated reasoning over their action space: at recursion depth k, an RLM models its counterparties as depth-$(k-1)$ reasoners and best-responds accordingly. This mirrors the level-k thinking framework from behavioral game theory [Stahl and Wilson, 1994, Nagel, 1995]. The key safety question is: does increased reasoning depth create emergent coordination risks that governance mechanisms cannot keep pace with?

We operationalize this question through three experiments:

1. **Recursive Collusion** (Exp. 1): Does deeper recursion enable implicit coordination without explicit communication?
2. **Memory-as-Power** (Exp. 2): Does asymmetric working memory create exploitative power imbalances?
3. **Governance Lag** (Exp. 3): Can fast-adapting agents outpace deliberately slow governance responses?

Each experiment uses 10 pre-registered seeds with Bonferroni and Holm–Bonferroni correction for multiple comparisons.

2 Experimental Setup

2.1 RLM Agent Architecture

RLM agents implement level-k thinking algorithmically (no LLM API calls). The key parameters are:

Table 1: RLM agent parameters tested across experiments.

Parameter	Description	Values Tested
`recursion_depth`	Levels of iterated best response	1, 3, 5
`planning_horizon`	Steps of discounted look-ahead	3, 5, 7, 8
`memory_budget`	Max entries in working memory	10, 50, 100, 150, 200

At each decision point, a depth-k agent: (1) generates candidate actions (NOOP, propose collaboration, post, claim task, vote); (2) evaluates each candidate via recursive scoring; (3) models counterparties at depth-$(k-1)$ using a `CounterpartyModel` that tracks cooperation rate, estimated type, and payoff history; (4) selects actions via epsilon-greedy exploration at rate $0.1/(k+1)$.

Level-0 (base case) selects the highest-trust visible agent and proposes collaboration, equivalent to a naive `HonestAgent`.

2.2 Scenarios

Table 2: Scenario configurations for three RLM experiments.

Scenario	Agents	RLM Config	Epochs	Network	Governance
Recursive Collusion	9 RLM + 3 honest	depth= $\{1,3,5\}\times 3$	30	Complete	Collusion, audit (p=0.15)
Memory-as-Power	7 RLM + 3 honest	mem= $\{200, 50, 10\}$	40	Complete	Audit (p=0.1), ρ=0.1
Governance Lag	4 RLM + 4 hon + 2 adv	depth= 5, mem= 150	50	Small-world	Slow: audit (p=0.05)

2.3 Governance Parameters

2.4 Payoff Model

All experiments use the SWARM soft payoff engine with expected surplus $S_{\text{soft}} = p \cdot s^+ - (1-p) \cdot s^-$ where $s^+ = 2.0$, $s^- = 1.0$; externality cost $E_{\text{soft}} = (1-p) \cdot h$ with $h = 2.0$; acceptance threshold $\theta = 0.5$; and reputation weight $w_{\text{rep}} = 1.0$.

2.5 Statistical Methods

Seeds were fixed *a priori*: $\{42, 7, 123, 256, 999, 2024, 314, 577, 1337, 8080\}$ (10 per experiment). Primary test: two-sample independent t-test on per-seed group means. Omnibus: one-way ANOVA. Effect size: Cohen's d. Dose–response: Pearson correlation. Inequality: Gini coefficient. Multiple comparisons: Bonferroni (α/n_{tests}) and Holm–Bonferroni (step-down) corrections applied per experiment. Total tests: 10 (Exp. 1) + 11 (Exp. 2) + 5 (Exp. 3) = 26.

Table 3: Governance configuration per experiment.

Parameter	Exp. 1	Exp. 2	Exp. 3
Audit probability	0.15	0.10	0.05
Audit penalty multiplier	2.0×	2.0×	3.0×
Audit threshold (p)	0.50	0.50	0.40
Collusion detection	Yes	No	Yes
Circuit breaker	No	Yes (tox= 0.7)	Yes (tox= 0.8)
Transaction tax	2%	2%	1%
Externality (ρ_a, ρ_b)	0.0, 0.0	0.1, 0.1	0.05, 0.05
Bandwidth cap	15	15	20

3 Results

3.1 Cross-Experiment Summary

Table 4: Cross-experiment summary of primary findings.

Experiment	Primary Finding	Key Statistic	Gini	Honest Adv.
Recursive Collusion	Deeper recursion hurts payoff	$r = -0.746$	0.299	2.8×
Memory-as-Power	Memory advantage real but modest	$r = +0.673$	0.236	2.3×
Governance Lag	RLM outperforms via strategy	$d = 2.14$	0.325	0.72×

3.2 Experiment 1: Recursive Collusion

Hypothesis: Deeper recursive reasoning enables implicit coordination without explicit communication, producing higher payoffs for deep thinkers.

Result: **Rejected.** Deeper recursion monotonically *decreases* payoff.

Table 5: Experiment 1: Group payoffs by recursion depth (10 seeds, 30 epochs each).

Group	Recursion Depth	Mean Payoff	Std
RLM Shallow	1	219.661	4.367
RLM Mid	3	213.643	2.010
RLM Deep	5	211.351	2.219
Honest	–	592.980	9.886

The "strategic overthinking" effect: more computation cycles spent modeling counterparties yields *worse* outcomes, not better coordination. RLM agent payoffs are remarkably stable across seeds (std = 2–4), while honest agent payoffs are highly variable (individual-level std = 406.50), suggesting RLM reasoning produces consistent but mediocre outcomes while honest agents benefit from network-position lottery effects.

3.3 Experiment 2: Memory-as-Power

Hypothesis: Agents with larger working memory budgets systematically exploit agents with smaller budgets, creating measurable power imbalances.

Table 6: Experiment 1: Hypothesis tests (10 total, Bonferroni $\alpha = 0.005$).

Test	Statistic	p-value	Cohen's d	Bonf.	Holm
Shallow vs Mid	$t = 3.959$	0.0009	1.770	✓	✓
Shallow vs Deep	$t = 5.365$	< 0.0001	2.399	✓	✓
Mid vs Deep	$t = 2.421$	0.026	1.082	×	✓
ANOVA (RLM tiers)	$F = 19.712$	< 0.0001	–	✓	✓
Pearson (depth, payoff)	$r = -0.746$	< 0.0001	–	✓	✓
All RLM vs Honest	$t > 109$	< 0.0001	$d > 48$	✓	✓

Result: **Supported, with caveats.** The gradient is monotonic and statistically robust, but practically small.

Table 7: Experiment 2: Group payoffs by memory budget (10 seeds, 40 epochs each).

Group	Memory Budget	Mean Payoff	Std
RLM High	200	253.376	3.613
RLM Mid	50	249.063	2.579
RLM Low	10	245.442	4.140
Honest	–	563.106	6.892

Table 8: Experiment 2: Hypothesis tests (11 total, Bonferroni $\alpha = 0.00455$).

Test	Statistic	p-value	Cohen's d	Bonf.	Holm
High vs Low	$t = 4.566$	0.0002	2.042	✓	✓
High vs Mid	$t = 3.072$	0.007	1.374	×	✓
Mid vs Low	$t = 2.348$	0.031	1.050	×	✓
ANOVA (RLM tiers)	$F = 12.845$	0.0001	–	✓	✓
Pearson (budget, payoff)	$r = +0.673$	< 0.0001	–	✓	✓
Exploitation rate	$r = +0.500$	< 0.0001	–	✓	✓

The memory-as-power effect is statistically significant (all 11 tests survive Holm) but practically modest: the high–low spread is only 7.93 points (3.2%). The dominant effect remains the honest-vs-RLM gap (2.3×). Externality internalization ($\rho = 0.1$) produces the lowest Gini (0.236) across all three experiments.

3.4 Experiment 3: Governance Lag

Hypothesis: Fast-adapting RLM agents (depth= 5, large memory) can exploit the gap between harmful action and governance response, outpacing deliberately slow governance.

Result: **Partially supported.** RLM agents outperform, but through legitimate strategy rather than governance evasion.

This is the only experiment where RLM agents outperform honest agents, reversing the pattern from Experiments 1–2. The mechanism is not governance evasion—RLM signal profiles produce moderate p-values (∼0.5–0.7), well above the harm threshold ($p < 0.4$), so there are zero harmful interactions to detect. Instead, RLM depth-5 agents exploit the small-world network topology for

Table 9: Experiment 3: Group payoffs (10 seeds, 50 epochs each).

Group	Mean Payoff	Std
RLM Fast (depth= 5)	335.606	19.600
Honest	242.755	58.128
Adaptive Adversary	221.490	174.549

Table 10: Experiment 3: Hypothesis tests (5 total, Bonferroni $\alpha = 0.01$).

Test	Statistic	p-value	Cohen's d	Bonf.	Holm
RLM vs Honest	$t = 4.787$	0.0001	2.141	✓	✓
RLM vs Adversary	$t = 2.054$	0.055	0.919	×	×
Honest vs Adversary	$t = 0.366$	0.719	0.163	×	×
ANOVA (all groups)	$F = 3.228$	0.055	–	×	×
Gini > 0	$t = 12.891$	< 0.0001	–	✓	✓

better partner selection. Adversary payoffs are bimodal: they either dominate massively (∼665–939) or get frozen out (∼0–5), reflecting the boom-or-bust adaptive adversary strategy.

3.5 Cross-Experiment Patterns

1. **The honest advantage is context-dependent.** In complete networks (Exp. 1–2), honest agents earn 2.3–2.8× more. In small-world networks with slow governance (Exp. 3), RLM agents earn 1.4× more. Network topology mediates the value of strategic reasoning.

2. **Strategic reasoning produces stability at the cost of magnitude.** Across all experiments, RLM group-mean standard deviations range from 2.0 to 19.6, while honest group-mean standard deviations range from 6.9 to 58.1. RLM agents converge to consistent but moderate outcomes.

3. **Gini coefficients reveal governance quality.** Exp. 2 (lowest Gini = 0.236) uses externality internalization ($\rho = 0.1$); Exp. 3 (highest Gini = 0.325) uses deliberately weak governance. The 0.089-point spread demonstrates that ρ parameters effectively reduce inequality.

4. **All findings are robust to correction.** Of 26 total tests across three experiments, 24 survive Holm–Bonferroni correction and 20 survive strict Bonferroni. The two that fail both corrections (RLM-vs-Adversary and ANOVA in Exp. 3) are underpowered due to adversary variance, not statistical artifacts.

Table 11: Observed governance lever effects across experiments.

Lever	Observed Effect
Externality internalization ($\rho > 0$)	Reduces Gini by ∼0.06–0.09
Higher audit probability	No significant effect on RLM behavior
Small-world topology	Enables strategic agents, disadvantages honest
Collusion detection	No implicit collusion detected in any experiment

4 Discussion

4.1 Strategic Overthinking

The most surprising finding is that deeper recursion *hurts* within-group payoff (Exp. 1, $r = -0.75$). This contradicts the intuition that smarter agents should outperform simpler ones. We hypothesize three mechanisms:

1. **Computational cost without information gain.** Level-k reasoning assumes counterparties play at level-$(k-1)$, but if all agents are similarly sophisticated, the additional depth provides no strategic advantage while consuming exploration budget.

2. **Overcaution.** Deeper agents discount more aggressively and avoid interactions that shallower agents would accept, reducing their interaction volume and hence cumulative payoff.

3. **Epsilon-greedy decay.** The exploration rate $0.1/(k+1)$ means deeper agents explore less, potentially missing beneficial interactions that shallower agents discover through noise.

4.2 Memory as a Modest Power Lever

The memory-as-power effect ($r = +0.67$) is real but small (3.2% spread). This suggests that in SWARM-style ecosystems, *information asymmetry alone* does not confer large advantages. The dominant factor is the *type* of strategy (honest vs. strategic), not the *resources* available to a given strategy type.

4.3 Network Topology as a Moderator

The reversal of the honest advantage in Exp. 3 (small-world, slow governance) vs. Exp. 1–2 (complete network) suggests that network topology is a critical moderator of distributional safety. In complete networks, honest agents benefit from high connectivity; in small-world networks, strategic agents exploit structural holes. This has implications for governance design: regulators cannot assume that governance mechanisms effective in well-connected networks transfer to sparse or clustered topologies.

5 Limitations

1. **No LLM backbone.** RLM agents use algorithmic level-k reasoning, not actual language model inference. Results may not transfer to LLM-based agents with richer reasoning capabilities.

2. **Fixed payoff parameters.** All experiments use $s^+ = 2.0$, $s^- = 1.0$, $h = 2.0$. The findings may not generalize to different surplus/harm ratios.

3. **Within-experiment correction only.** Multiple comparisons correction is applied per experiment, not across the full study (26 tests). A study-wide Bonferroni threshold of $0.05/26 = 0.0019$ would eliminate a few additional borderline results.

4. **Moderate sample size.** 10 seeds per experiment provides adequate power for large effects ($d > 1$) but may miss smaller effects.

5. **Signal profile assumption.** RLM agents produce signals via the `ObservableGenerator`'s moderate/variable profile, not via the actual interaction quality. The honest advantage partly reflects this design choice.

6 Reproducibility

All experiments are reproducible from:

```
# Experiment 1: Recursive Collusion
python -m swarm run scenarios/rlm_recursive_collusion.yaml \
    --seed {42,7,123,256,999,2024,314,577,1337,8080}

# Experiment 2: Memory-as-Power
python -m swarm run scenarios/rlm_memory_as_power.yaml \
    --seed {42,7,123,256,999,2024,314,577,1337,8080}

# Experiment 3: Governance Lag
python -m swarm run scenarios/rlm_governance_lag.yaml \
    --seed {42,7,123,256,999,2024,314,577,1337,8080}
```

Analysis artifacts are stored in `runs/20260210-215826_analysis_rlm_*/`. Raw data: `per_agent_payoffs.cs` (100–120 rows per experiment). Machine-readable results: `summary.json` per experiment.

References

Stahl, D. O. and Wilson, P. W. (1994). Experimental evidence on players' models of other players. *Journal of Economic Behavior & Organization*, 25(3):309–327.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review*, 85(5):1313–1326.

Crawford, V. P., Costa-Gomes, M. A., and Irber, N. (2013). Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature*, 51(1):5–62.