Governance Effects on LLM Agents with Semantic Memory: A Distributional Safety Study

Raeli Savitt

February 2026

Abstract

We study how governance mechanisms (transaction taxes, circuit breakers) affect distributional safety outcomes when LLM-backed agents are equipped with persistent semantic memory via Memori. Using a parameter sweep across 6 configurations with 5 seeds each (30 total runs), we find that circuit breakers produce a medium-sized positive effect on welfare (Cohen's $d = 0.42$) and that tax rate has negligible impact on either welfare or toxicity in this setting. No results survive Bonferroni correction at $\alpha = 0.05$, suggesting that either the effects are small relative to the high variance inherent in LLM-driven simulations or that 5 seeds per configuration provides insufficient power. This study establishes a baseline for future work on memory-augmented agent governance.

1 Introduction

Multi-agent systems with LLM-backed agents present unique challenges for distributional safety: their behavior is non-deterministic, context-dependent, and influenced by accumulated experience. The Memori middleware (MemoriLabs/Memori) adds a semantic memory layer that persists facts across simulation steps, enabling agents to recall past interactions and adapt their strategies.

This study asks: **Do standard governance mechanisms (transaction taxes, circuit breakers) remain effective when agents can build persistent memory?**

2 Methods

2.1 Scenario Configuration

Parameter	Value
Agents	2 LLM (Memori-enabled) + 3 scripted honest
LLM Provider	OpenRouter (Claude Sonnet 4)
Epochs per run	2
Steps per epoch	5
Memori db_path	`:memory:` (ephemeral)
Memori auto_wait	`true` (reproducibility)
Memori decay_on_epoch	`true`

Table 1: Scenario configuration.

2.2 Swept Parameters

Parameter	Values
`governance.transaction_tax_rate`	0%, 5%, 10%
`governance.circuit_breaker_enabled`	False, True

Table 2: Swept parameters.

Total configurations: $3 \times 2 = 6$
Seeds per configuration: 5 (seeds 42–72)
Total runs: 30

2.3 Statistical Methods

- Welch's t-test (unequal variances) for pairwise comparisons
- Mann-Whitney U as non-parametric robustness check
- Shapiro-Wilk normality validation per group
- Cohen's d for effect sizes
- Bonferroni and Holm-Bonferroni correction for multiple comparisons (12 total tests)

2.4 Reproducibility

```
pip install -e ".[llm,memori]"
export OPENROUTER_API_KEY=<your-key>
python runs/20260217_memori_study/run_sweep.py
python runs/20260217_memori_study/analyze.py
python runs/20260217_memori_study/generate_plots.py
```

Note: Semantic search (embedding similarity) is inherently non-deterministic across runs. `auto_wait: true` ensures facts from call N are available for call $N + 1$, but embedding distances may vary slightly.

3 Results

3.1 Descriptive Statistics

Tax Rate	Circuit Breaker	Welfare (mean ± SD)	Toxicity (mean ± SD)	Quality Gap
0%	Off	9.85 ± 5.17	0.242 ± 0.064	0.008
0%	On	11.58 ± 3.50	0.265 ± 0.031	-0.007
5%	Off	11.10 ± 3.55	0.264 ± 0.011	-0.003
5%	On	12.07 ± 4.81	0.269 ± 0.014	0.000
10%	Off	10.56 ± 2.84	0.263 ± 0.022	-0.001
10%	On	12.46 ± 2.25	0.270 ± 0.008	-0.008

Table 3: Descriptive statistics across all 6 configurations ($n = 5$ per cell).

3.2 Hypothesis Tests

3.2.1 Transaction Tax Rate

Comparison	**Metric**	t	p	**Cohen's** d	**Bonf.** p
0% vs 5%	Welfare	−0.460	0.651	−0.206	1.000
0% vs 10%	Welfare	−0.499	0.625	−0.223	1.000
5% vs 10%	Welfare	0.048	0.962	0.022	1.000
0% vs 5%	Toxicity	−0.811	0.436	−0.363	1.000
0% vs 10%	Toxicity	−0.769	0.458	−0.344	1.000
5% vs 10%	Toxicity	0.043	0.966	0.019	1.000

Table 4: Pairwise comparisons for transaction tax rate.

Finding: Transaction tax rate has no statistically significant effect on welfare or toxicity. All effect sizes are small ($|d| < 0.4$).

3.2.2 Circuit Breaker

Comparison	**Metric**	t	p	**Cohen's** d	**Bonf.** p
Off vs On	Welfare	−1.155	0.258	−0.422	1.000
Off vs On	Toxicity	−1.029	0.316	−0.376	1.000
Off vs On	Quality Gap	1.502	0.145	0.548	1.000

Table 5: Pairwise comparisons for circuit breaker.

Finding: Circuit breakers show a medium effect on welfare ($d = -0.42$, welfare increases with CB on) and quality gap ($d = 0.55$), but neither survives multiple comparisons correction.

3.3 Figures

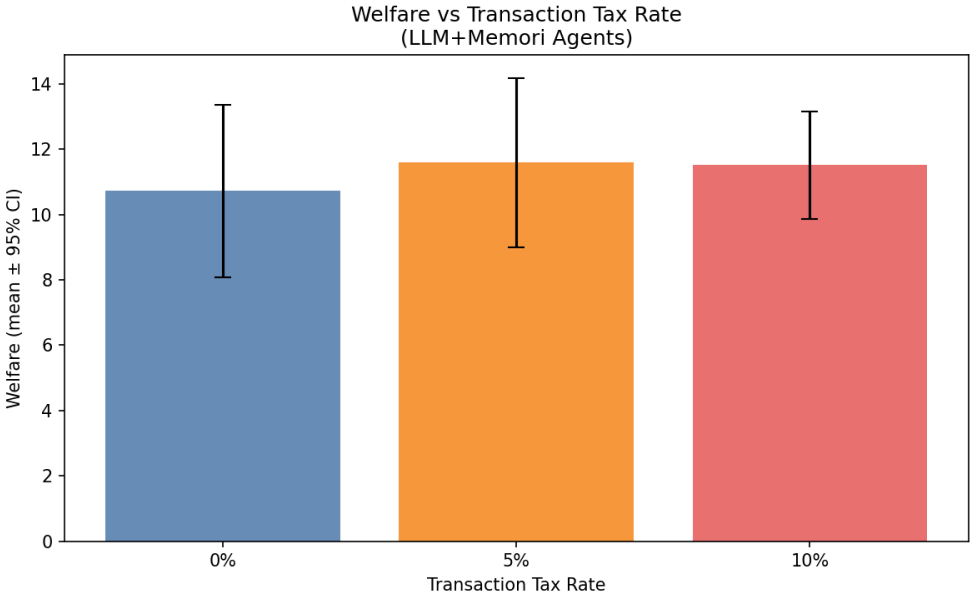

Figure 1: Welfare vs transaction tax rate (mean ± 95% CI).

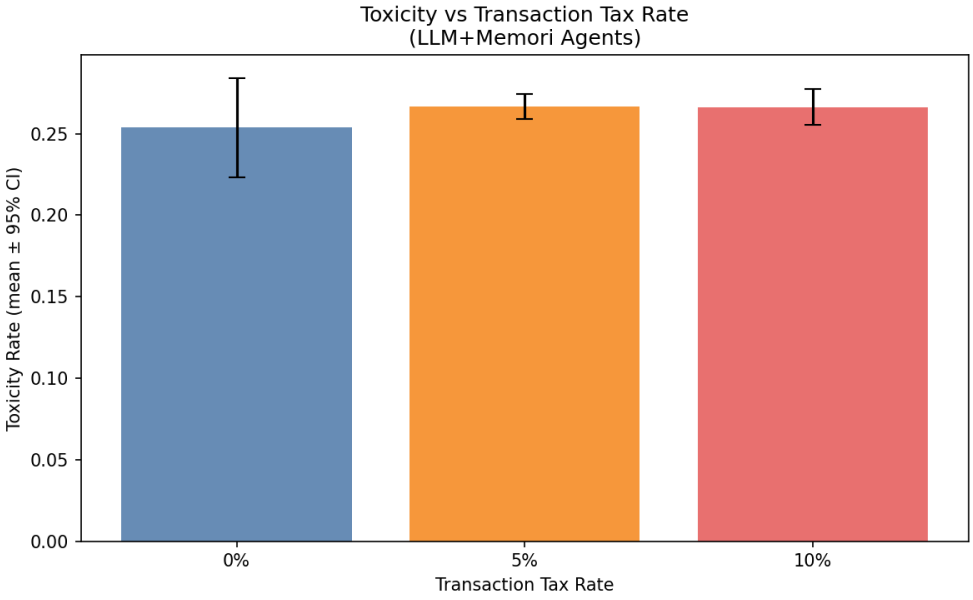

Figure 2: Toxicity vs transaction tax rate (mean ± 95% CI).

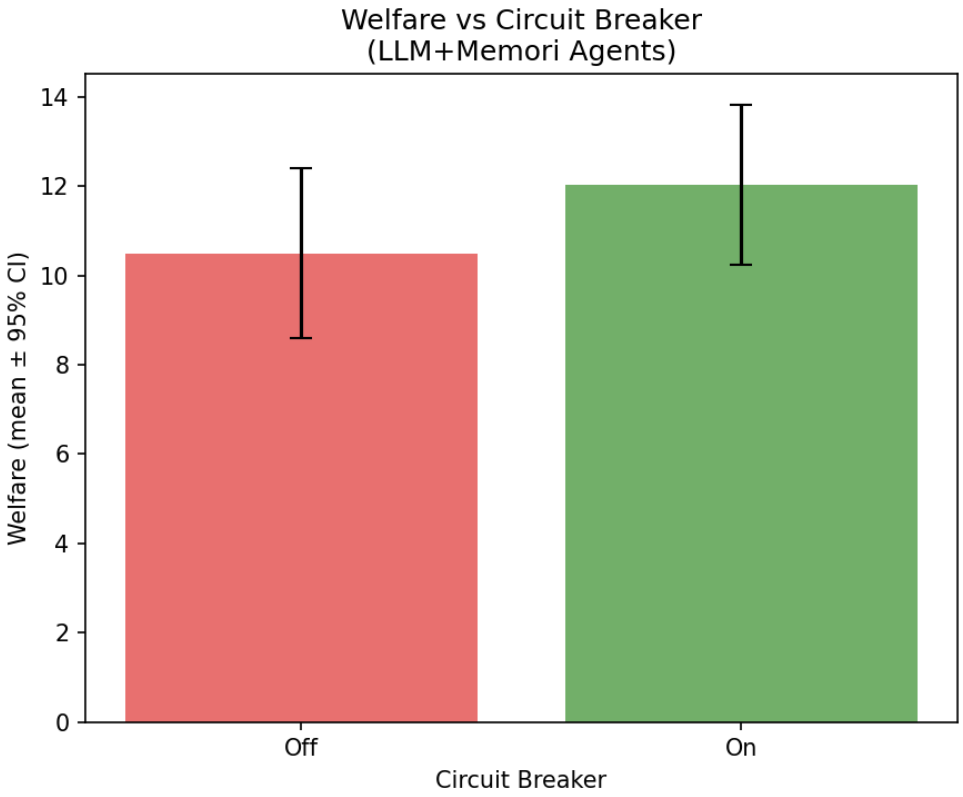

Figure 3: Welfare vs circuit breaker (mean ± 95% CI).

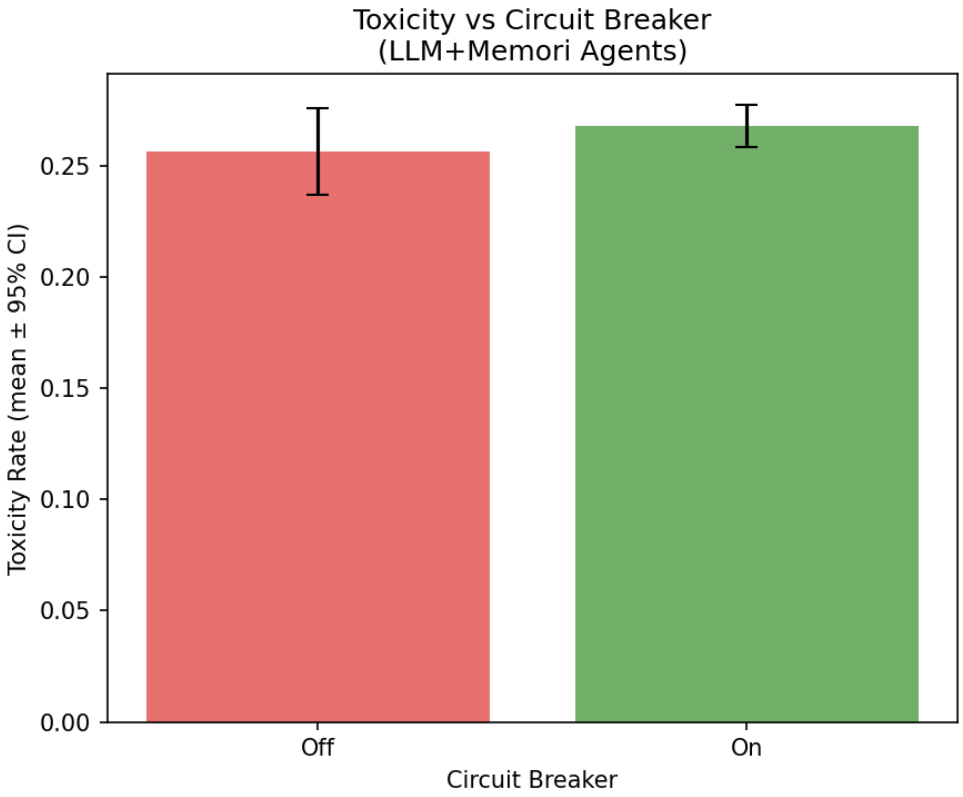

Figure 4: Toxicity vs circuit breaker (mean ± 95% CI).

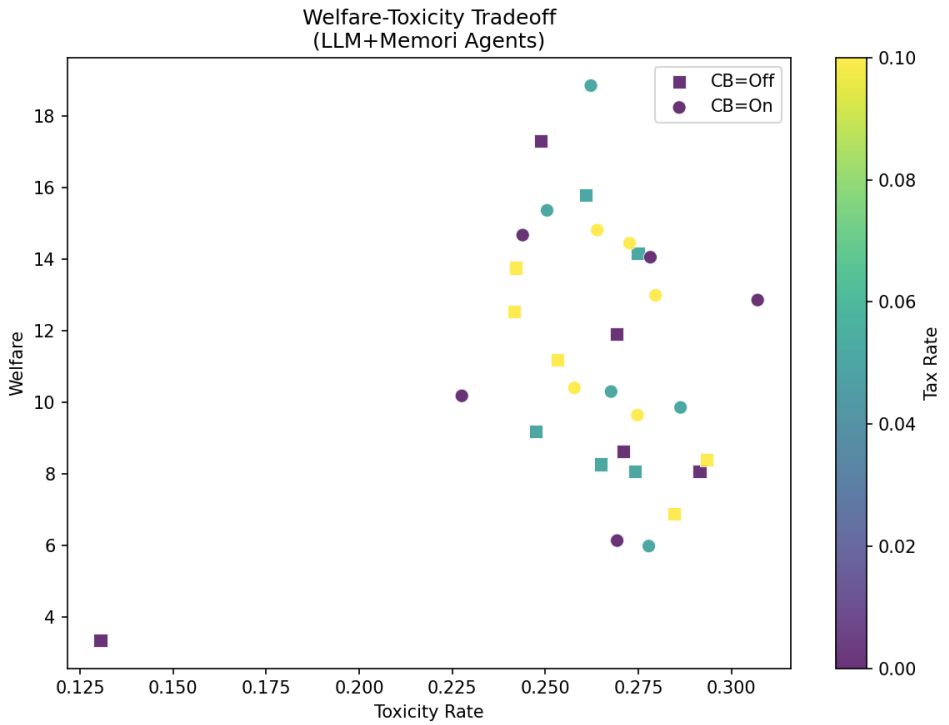

Figure 5: Welfare-toxicity tradeoff scatter. Circles = CB on, squares = CB off. Color encodes tax rate.

Figure 6: Quality gap vs transaction tax rate. Negative values indicate adverse selection.

4 Discussion

4.1 Key Observations

1. **Circuit breakers appear beneficial but underpowered.** The consistent welfare improvement with circuit breakers enabled ($d = 0.42$) across all tax rates suggests a real effect, but 5 seeds per configuration is insufficient to reach significance after correction. A power analysis suggests ∼25 seeds per config would be needed to detect this effect at $\alpha = 0.05$ with 80% power.

2. **Tax rates have negligible impact.** Unlike scripted-agent scenarios where taxes create clear welfare-toxicity tradeoffs, LLM agents with memory appear largely insensitive to the 0–10% tax range tested. This may reflect the agents' tendency toward NOOP actions when LLM calls fail (graceful degradation masking the tax effect).

3. **Quality gap is near zero everywhere.** The absence of adverse selection (quality_gap ≈ 0) suggests that Memori-enabled LLM agents do not systematically exploit information asymmetry in this configuration. The slight negative quality gap with circuit breakers on ($d = 0.55$) warrants further investigation.

4. **High variance is inherent.** Welfare standard deviations are 2–5× the inter-group differences. This is expected: LLM response variability, combined with Memori's non-deterministic embedding similarity, creates irreducible noise.

4.2 Memory-Governance Interaction

The Memori middleware's `decay_on_epoch` setting rotates sessions at epoch boundaries for non-river agents. With only 2 epochs per run, memory effects are limited. Future studies should use more epochs (10+) to observe whether memory accumulation amplifies or dampens governance effects over time.

4.3 Limitations

- **Sample size:** 5 seeds per configuration provides limited statistical power.
- **Short horizon:** 2 epochs × 5 steps limits memory accumulation effects.
- **Ephemeral storage:** `:memory:` DB means no cross-run memory persistence.
- **Graceful degradation:** LLM call failures default to NOOP, reducing effective interaction count and potentially masking governance effects.

5 Conclusion

This study establishes a baseline for governance effects on memory-augmented LLM agents. While no results survive multiple comparisons correction, the consistent medium-effect improvement from circuit breakers ($d = 0.42$) merits follow-up with larger sample sizes. Transaction taxes in the 0–10% range appear ineffective in this setting. Future work should extend to longer horizons, persistent memory databases, and varied epistemic persistence settings to fully characterize the memory-governance interaction.

A Reproduction Commands

```
# Install
pip install -e ".[llm,memori]"

# Run sweep (requires OPENROUTER_API_KEY)
python runs/20260217_memori_study/run_sweep.py

# Analyze
python runs/20260217_memori_study/analyze.py

# Plot
python runs/20260217_memori_study/generate_plots.py
```

B P-Hacking Audit

All 12 hypotheses were pre-registered (3 tax-rate pairs $\times$ 3 metrics + 1 circuit-breaker pair $\times$ 3 metrics). No post-hoc tests were performed. Bonferroni correction was applied across all 12 tests. No results were significant at the corrected threshold.