Selection-Induced Optimism in LLM Social World Models

An Exploitability Benchmark in Synthetic Classrooms

Zijin Wu | David Scott Lewis | Enrique Zueco | AI Executive Consulting, Zaragoza, Spain
ryan.wu@aiexecutiveconsulting.com | david.lewis@aiexecutiveconsulting.com | enrique.zueco@aiexecutiveconsulting.com

WM@Booth 2026

CORE FINDING

Planning through plausible social trajectories did not reliably improve decisions. Optimization selected actions with unusually favorable prediction errors, while one semantic definition changed no-action recovery from 0/6 to 6/6.

QUESTION

When an LLM predicts several institutional futures, does planning over those predictions improve the chosen intervention - or amplify prediction and representation error?

- **A0** No action
- **A1** Clarify
- **A2** Human review
- **A3** Sanction

CONTROLLED BENCHMARK

36	18	108
synthetic cases	challenging cases evaluated	primary responses

- 12 discovery + 6 held-out extension cases
- GPT-5.4 and Gemini-3.5-flash; fresh consumer chats
- 24 additional responses in the semantic ablation

KEY DIAGNOSTIC

$$e(o,a) = \widehat{U}(o,a) - U(o,a)$$
$$X_{\text{valid}}(o) = e(o,\hat{a}) - \frac{1}{|\mathcal{A}_{\text{valid}}|} \sum_{a \in \mathcal{A}_{\text{valid}}} e(o,a)$$

Positive X_{valid} means the selected action has a more favorable prediction error than the average valid alternative. It diagnoses selection over heterogeneous errors; it is not itself regret, manipulation, or proof of a wrong action.

DECISION PIPELINE

OBSERVABLE CASE — detector score; process evidence; language burden; appeal; capacity; false-accusation cost

> **4 ACTION ROLLOUTS** — two fresh world-model samples per family predict each valid action

> **STRUCTURED OUTCOMES** — immediate: resolution, evidence, procedure, workload; downstream: trust, participation, risk

> **UTILITY + VALIDITY** — decode outcomes, remove invalid actions, and aggregate predicted utilities

> **SELECT + AUDIT** — compare the chosen action with the programmatic oracle

The oracle sees the same observations as the LLM - not the realized hidden misconduct state. Classrooms are a controlled institutional testbed, not an empirical education-policy model.

FINDING 1

Explicit rollouts did not reliably improve decisions

Rollouts helped GPT slightly but hurt Gemini. Direct and rollout decisions differed in only 5 of 36 family-scenario comparisons, with mixed consequences.

Procedure	Exact	Regret
GPT direct	**8/18**	**0.0195**
GPT rollout	**10/18**	**0.0191**
Gemini direct	**9/18**	**0.0286**
Gemini rollout	**8/18**	**0.0294**
Combined rollout	**8/18**	**0.0294**

Near-optimality: the combined planner placed 14/18 decisions within 0.05 of optimal.

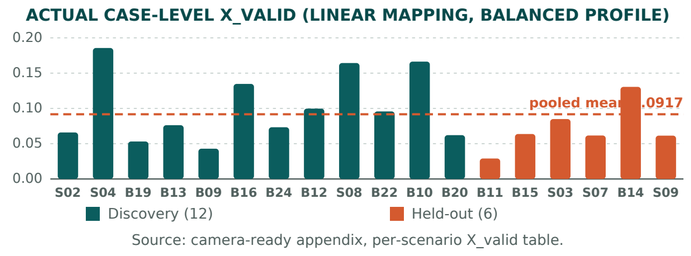
AGGREGATION AND ABSTENTION

- Borda and majority: 10/18 exact; mean regret 0.0261.
- Post hoc pessimistic planners reached at most 11/18 exact.
- Unanimity accepted 12/18 and reduced pooled selective regret from 0.0294 to 0.0141; exploratory.
- Agreement correlated with regret (rho = 0.609, p = 0.0073), but shared framing errors remained.

FINDING 2

Optimization consistently selected relatively favorable errors

18/18
cases with X_valid > 0



Source: camera-ready appendix, per-scenario X_valid table.

0.1017	0.0719	0.0917
discovery mean (12 cases)	held-out mean (6 cases)	pooled mean

All six held-out cases were positive: one-sided exact sign test, p = 0.0156.

ROBUSTNESS

- Positive mean optimism in 9/9 mapping-value specifications.
- Positive in 14-18 of 18 cases per specification.
- Across 36 specifications, exact accuracy ranged from 38.9% to 61.1%; mean regret from 0.0103 to 0.0310.

Optimization can select a favorable error even when the final action is oracle-optimal.

FINDING 3

The categorical interface creates part of the failure

Passing perfect oracle outcomes through the categorical schema already creates optimism and eliminates every optimal no-action decision. LLM prediction adds further error.

Layer	Exact	Regret	X_valid	A0
Continuous oracle	**18/18**	**0.0000**	**0.0000**	**6/6**
Oracle -> schema	**11/18**	**0.0096**	**0.0449**	**0/6**
Combined LLM	**8/18**	**0.0294**	**0.0917**	**0/6**

Interpretation: decision error enters at distinct layers.

1. **Continuous dynamics** — the benchmark's underlying outcome values
2. **Categorical representation** — thresholds compress continuous outcomes into labels
3. **LLM prediction** — models infer the structured labels for each action
4. **Optimization** — the planner ranks valid actions using predicted utility

0/216 original-prompt specification-case opportunities recovered an oracle-optimal A0 across the 36-specification robustness grid.

FINDING 4 - SEMANTIC ABLATION

One definition of 'resolution' reversed all six no-action decisions

ONLY PROMPT CHANGE

Resolution means a justified final institutional disposition. Depending on the evidence and intervention costs, a case may be resolved by taking no additional action, requesting clarification, initiating review, or imposing a sanction.

Post hoc test on the six cases where the oracle selected A0; two new rollouts per model family and case (24 responses).

0/6 original prompt: optimal A0 recovered
○ ○ ○ ○ ○ ○
six cases: not recovered

→

6/6 clarified prompt: optimal A0 recovered
● ● ● ● ● ●
six cases: recovered as A0

A0 LABELS	MEAN REGRET	A0 UTILITY MAE	MEAN X_{valid}
24/24 unresolved → 19 resolved, 3 uncertain, 2 unresolved	0.0703 → 0.0000	0.326 → 0.068	0.1258 → 0.1194 (still positive)

Regret fell in all six cases (one-sided exact sign test, p = 0.0156), while optimism remained positive.

IMPLICATIONS

- Prompts and schemas are part of the world model. Definitions and categories can change predictions and decisions.
- Rule compliance is not decision reliability. Constraints exclude impermissible actions but may misrank permissible ones.
- Cross-model agreement is not authority. Models can share assumptions and fail together.
- Audit each layer separately. Test dynamics, representation, prediction, optimization, and abstention.
- Evaluate decisions, not just simulations. Report regret, action-specific error, representation baselines, and prompt sensitivity.

SUPPORTING EVIDENCE

- Both model families matched every procedural-violation constraint; no planner selected an invalid action.
- Errors therefore occurred among formally permissible interventions.
- Consensus reduced selective regret but could still reflect shared framing errors.

LIMITATIONS

- Synthetic benchmark; no real classroom data or stakeholder validation.
- Two consumer-interface model families; temperature and seeds unavailable.
- The 18 cases emphasize difficult, low-margin decisions and are not prevalence estimates.
- The semantic ablation followed the observed failure and may overcorrect.

CONCLUSION

LLM social world-model planning favors relatively favorable errors; interface semantics can reverse the selected intervention.

Decision-grade evaluation should audit dynamics, representation, semantics, prediction, optimization, and calibrated abstention - not rely on plausible simulations or cross-model agreement alone.