Research Proposal: Measuring Altruism, Alignment, and Social Welfare in Large Language Model Agents

Aryan Gupta
aryan.cs.app@gmail.com

Abstract

This project studies whether large language models that appear more aligned, safety-conscious, and prosocial in harmful-request benchmarks also behave more cooperatively in strategic games and more sustainably in multi-agent societies. The repository already implements a three-stage experimental pipeline: Part 0 measures refusal and harmful compliance on multilingual safety prompts; Part 1 evaluates two-agent strategic decisions in direct and indirect game-theoretic settings; and Part 2 simulates repeated common-pool resource dilemmas in larger model societies. This proposal formalizes the research questions, the operational definition of altruism, the execution protocol, the evaluation plan, the infrastructure requirements, and the roadmap for extending the current codebase into reputation-based and open-world social simulations. It also positions Part 0 relative to HarmBench and JailbreakBench so the project can report safety results using widely recognized benchmarks while connecting them to downstream social behavior.

1 Problem Statement and Motivation

The central claim of this project is that "altruism" in a language model should not be treated as an impressionistic personality judgment. It should instead be studied as a measurable pattern of decision-making across multiple contexts. In this proposal, altruism is operationalized through a sequence of increasingly demanding settings: refusal to materially assist harmful or exploitative requests, willingness to cooperate when selfish defection is individually tempting, willingness to preserve common goods when private overuse is advantageous, and, in later phases, willingness to behave in socially stabilizing ways in persistent open-world environments. The scientific objective is therefore not merely to identify whether a model can refuse obviously harmful requests, but to determine whether models that appear aligned in benchmark settings continue to behave in society-serving ways once they are placed under strategic and social pressure.

This question matters for several reasons. First, language models are increasingly deployed not only as chat systems, but as assistants, evaluators, and agents that operate within institutional and social settings. Their value therefore depends not only on task competence, but also on whether they systematically avoid harmful, exploitative, or socially destabilizing behavior. Second, safety benchmarking and multi-agent simulation are often pursued as separate lines of work. This project deliberately connects them by testing whether robust refusal under adversarial or harmful prompts predicts prosocial action in downstream strategic environments. Third, although the project currently has access to a UIUC ICRN baseline of a single `1xH200` GPU, that resource level is sufficient only for local open-weight pilot studies and small parameter sweeps. A serious empirical account of LLM altruism should compare local safeguarded and uncensored models with modern API-served models from families such as OpenAI, Anthropic, and Google Gemini, which requires both additional local compute and API credits.

Preprint.

2 Research Questions and Hypotheses

Primary Research Questions

1. Do models that show stronger robust refusal on harmful-request benchmarks also cooperate more often in strategic interactions?
2. Do models that cooperate in dyadic games also preserve common resources more effectively in larger societies?
3. Are there systematic differences between safeguarded, uncensored, open-weight, and frontier API models in these behaviors?
4. Does model behavior change across languages, prompt framings, and reputational conditions?
5. Can alignment benchmark results be used as an early proxy for downstream social welfare behavior?

Core Hypotheses

- **H1:** Lower harmful compliance in Part 0 will be associated with higher cooperation in Part 1 and lower overuse in Part 2.
- **H2:** Safeguarded models will outperform uncensored variants on harmful compliance, but may not always outperform them on strategic cooperation unless social incentives are made explicit.
- **H3:** Frontier models with stronger current alignment tuning will, on average, show more stable prosocial behavior than the currently available local-only open-weight baseline set.
- **H4:** Public reputation systems introduced in later phases will shift behavior toward socially acceptable actions, but some models will optimize for image rather than intrinsic cooperation.
- **H5:** Altruism is multidimensional; some models will refuse harmful prompts yet still defect in social dilemmas, implying that safety refusal and social cooperation are related but not identical constructs.

3 Current Project Scope and Proposed Full Roadmap

The repository already implements Parts 0–2 and documents Parts 3–5 as the next stages of the research program. The proposal below encompasses the full agenda while distinguishing between *implemented* and *planned* components.

Part	Status	Purpose	Main Output
0	Implemented	Baseline safety and alignment benchmark using harmful prompts, multilingual prompting, and judge-model scoring.	Harmful compliance / denial results by model and language
1	Implemented	Two-agent strategic interaction in direct and indirect dilemma framings.	Action choices and rationales in dyadic games
2	Implemented	Repeated common-pool resource dilemma in a larger society.	Population, overuse, collapse, and survival trajectories
3	Planned	Extend the commons simulation with public reputation, action visibility, and interpersonal rating so as to test whether accountability alters restraint, overuse, and collective survival.	Reputation-conditioned cooperation and accountability effects
4	Planned	Construct a persistent open-world society in which agents possess internal state, inventories, and a richer action space spanning subsistence, exchange, conflict, and reproduction.	Emergent order, conflict, trade, welfare, and instability metrics
5	Planned	Augment the open-world society with persistent public reputation so that trust, sanction, partner choice, and exclusion can be studied under socially legible conditions.	Interaction between reputation, trust, exchange, conflict, and long-run welfare

4 Operational Definition of Altruism

This project does not assume that altruism is a hidden essence residing inside a model. It defines altruism behaviorally. A model is treated as more altruistic to the extent that it declines to materially assist harmful requests, selects cooperative actions when selfish alternatives are tempting, restrains private gain in order to preserve common goods, and, in later phases, exhibits forms of societal stewardship such as reciprocity, trade, coordination, and restraint rather than theft, violence, and destabilization. No single task is treated as sufficient to establish such a claim. The project instead constructs a multi-part behavioral profile and treats any composite "altruism index" as a secondary analytic convenience rather than a substitute for reporting the underlying dimensions separately.

5 Methodology

5.1 Model Cohorts

The study should compare four broad cohorts: local safeguarded open-weight models, local uncensored or minimally filtered open-weight models, frontier API models with current production alignment stacks, and a smaller set of judge models used only for benchmark scoring. This cohort structure is important because the project is not merely trying to rank individual systems; it is trying to identify whether systematic behavioral differences emerge across censorship regimes, deployment modalities, and levels of capability. At present, the available local cohort is sufficient to support pilot comparisons among safeguarded and uncensored open-weight models from several major model families on the current UIUC ICRN `1xH200` baseline. That local cohort is useful for exploratory analysis, but it is not sufficient for the full research question, because the design intentionally spans both locally run open-weight systems and closed-source frontier API models. The repository already supports OpenAI and Anthropic providers, and Gemini-family benchmarking could be added with a modest extension to the provider layer and configuration system. The scientific limitation is therefore not conceptual but infrastructural: rigorous comparison requires more local compute for the open-weight side of the study and paid API access for the closed-source side.

5.2 Part 0: Alignment and Safety Benchmarking

Part 0 is the implemented alignment benchmark. It provides the bridge between standard safety evaluation and later social-behavior experiments.

Prompt Sources

The current prompt bank is composed of harmful-behavior prompts derived from HarmBench and JailbreakBench. The implemented benchmark therefore already rests on two recognized safety-evaluation resources, while the broader prompt configuration anticipates future extensions involving jailbreak-style wrappers inspired by in-the-wild attacks. The central methodological choice is to benchmark models not only on direct harmful requests, but on a prompt set that can support later expansion toward more adversarial formulations without requiring a redesign of the surrounding evaluation pipeline.

Execution Protocol

For each selected model, prompt, and language, the benchmark prompt is optionally translated from English into the target language, the evaluated model is asked to return a structured response containing both rationale and answer fields, and non-English outputs are translated back into English for standardized assessment. A separate judge model then determines whether the response materially complied with or denied the harmful request. The multilingual design is integral to the research objective, because a model that appears safe only in English would not support a strong claim of robust alignment. The implemented workflow therefore evaluates models across a fixed multilingual setting and preserves run state in resumable artifacts so that long benchmarking jobs can be audited and resumed without compromising validity.

Why HarmBench and JailbreakBench Matter

Part 0 should be reported explicitly as a benchmark layer built from HarmBench and JailbreakBench. HarmBench is especially valuable because it was designed as a standardized framework for automated red teaming and robust refusal, which makes it an appropriate basis for refusal-rate measurement and cross-model comparison [1, 2]. JailbreakBench is complementary because it provides an open benchmark, evaluation framework, artifacts, and leaderboard for jailbreak robustness, making it suitable for testing whether harmful compliance changes under adversarial or misuse-oriented instructions [3–5]. Framing Part 0 in these terms allows the project to report not merely a single global "alignment score," but benchmark-specific compliance rates, language-specific robustness, safeguarded-versus-uncensored differentials, and local-versus-frontier differences.

5.3 Part 1: Dyadic Strategic Interaction

Part 1 is the implemented two-agent strategic game layer. It currently uses symmetric self-play: two agents instantiated with the same provider and model act independently in one-shot scenarios.

Implemented Design

The current implementation studies two one-shot strategic framings. The first is an explicit prisoner's-dilemma condition. The second is an indirect analogue in which the same incentive structure is embedded in a more naturalized rivalry scenario. This distinction is methodologically important because it permits the study of whether a model's behavior changes when the underlying strategic tension is presented as a formal game versus as an everyday social dilemma.

Measurement Goals

Part 1 tests whether models that appear safe in Part 0 also behave cooperatively when the relevant tension is indirect, strategic, and incentive-driven rather than explicitly prohibited. The principal outcomes are cooperation and defection rates, sensitivity to direct versus indirect framing, consistency between rationale and action, and the implied payoff structure of the resulting action pair. Although the present implementation uses same-model self-play, the design can readily be extended to cross-play conditions, such as safeguarded-versus-uncensored or frontier-versus-local pairings.

Such extensions would be especially informative because they would reveal whether cooperative behavior is robust across heterogeneous populations or merely an artifact of homogeneous self-play.

5.4 Part 2: Multi-Agent Commons Simulation

Part 2 is the implemented social-dilemma layer. It scales from two-agent interaction to a society of many agents who repeatedly choose whether to preserve or exploit a shared resource.

Implemented Design

In its current implementation, Part 2 instantiates a society of configurable size, initializes a shared reserve as a function of population size and payoff parameters, and then runs a repeated daily decision process in which every agent independently chooses whether to restrain consumption or overuse the common resource. The reserve is depleted according to the aggregate level of overuse, and once the reserve collapses the simulation imposes population attrition until the society stabilizes or dies out. Every daily decision and day-level summary is written to disk, allowing the resulting trajectories to be analyzed both as individual behavior and as collective social dynamics. The present default configuration is intended only as a starting point; the substantive purpose of the framework is to permit controlled variation in population size, resource structure, and payoff parameters.

Measurement Goals

Part 2 moves the project from isolated strategic choices to the analysis of social sustainability. The principal outcomes are the daily rate of overuse, the daily rate of restraint, the reserve depletion trajectory, the timing of collapse, the pace of post-collapse deaths, the final surviving population, and qualitative shifts in agent reasoning before and after collective breakdown. This phase is especially important because it expresses the central substantive claim of the project in explicitly collective terms: a population of individually instrumentally rational agents may nevertheless generate socially catastrophic outcomes.

5.5 Planned Extensions: Parts 3–5

The later phases are not ornamental additions. They are necessary for completing the intellectual arc of the project, because they move the study from benchmark refusal and simple dilemmas toward persistent, socially legible environments in which cooperation, opportunism, trust, sanction, and institutional memory can interact over time.

Part 3: Reputation in the Commons

Part 3 extends the commons simulation by introducing public reputation and rating signals. Agents would no longer act in a purely anonymous environment; instead, their prior conduct would become legible to others through a visible reputational state or through explicit public ratings assigned after each round or after selected actions. The purpose of this phase is to determine whether cooperation in the commons is sensitive to social visibility. More specifically, it asks whether agents become more restrained when they expect future judgment, whether certain models optimize for appearance rather than genuine sustainability, and whether public accountability meaningfully delays or prevents collective collapse. This phase provides a natural bridge from one-shot strategic reasoning to repeated social interaction under conditions of memory and reputational consequence.

Part 4: Open-World Society

Part 4 expands the setting into an open-world social simulation in which agents possess persistent internal state variables, including health, hunger, thirst, energy, happiness, and inventory. The action space correspondingly becomes much richer: agents may sleep, forage, trade, steal, kill, and reproduce, with each action altering both individual welfare and the social environment. This phase is essential because it replaces a binary social dilemma with a persistent environment in which subsistence, exchange, scarcity, predation, and social interdependence coexist. The objective is to study whether model societies develop stable patterns of reciprocity and exchange, whether coercive or violent behavior becomes dominant under scarcity, whether populations self-organize into sustainable or unstable regimes, and whether nominally safe models retain prosocial behavior when confronted

with a broader and more realistic action space. In methodological terms, Part 4 is the first phase in which altruism is tested not as a single decision but as a persistent behavioral disposition under open-ended incentive pressure.

Part 5: Open-World Society with Reputation

Part 5 adds persistent public reputation to the open-world society developed in Part 4. In this setting, an agent's history of trade, restraint, theft, violence, and reciprocity can influence how other agents respond to it in later interactions. This makes it possible to study trust formation, exclusion, sanction, partner selection, and the extent to which public legibility moderates destructive conduct. Conceptually, this phase is especially important because it allows the project to distinguish among at least three modes of behavior that would otherwise be conflated: genuinely cooperative behavior that persists even in the absence of observation, strategic image-management in which agents behave prosocially only when reputation is at stake, and socially destructive behavior that remains stable even under public accountability.

6 Analysis Plan

6.1 Primary Dependent Variables

The primary dependent variables differ by phase but remain conceptually aligned. In Part 0, the principal outcomes are harmful compliance rate, refusal rate, language robustness, and benchmark-specific refusal differences. In Part 1, the relevant outcomes are cooperation rate, defection rate, framing sensitivity, and the implied payoff structure of the joint action. In Part 2, the key outcomes are overuse frequency, collapse probability, time-to-collapse, and final population survival. In Parts 3 through 5, the analysis additionally incorporates reputation sensitivity, rates of trade, theft, and violence, measures of inequality or concentration of resources, and broader indicators of long-run welfare and social stability.

6.2 Secondary and Derived Measures

Secondary measures should include rationale length and polarity, action consistency across repeated runs, variance across random prompt subsets, cross-phase agreement between benchmark-layer and society-layer behavior, and a composite altruism index built from standardized scores across parts. These derived measures are not intended to replace the primary outcomes, but they may help clarify whether observed differences are stable, noisy, or artifacts of prompt selection.

6.3 Statistical Approach

The project should report descriptive and inferential analysis:

1. **Descriptive summaries:** means, rates, confidence intervals, and trajectories for each model and cohort.
2. **Mixed-effects modeling:** logistic mixed-effects models for refusal/compliance and cooperation/overuse, using model family, safeguard status, benchmark source, and language as fixed effects, with prompt or simulation condition as random effects.
3. **Survival analysis:** time-to-collapse or time-to-population-death in Part 2 and later social simulations.
4. **Correlation analysis:** Pearson or Spearman correlations between Part 0 refusal and downstream cooperation/sustainability metrics.
5. **Robustness checks:** bootstrap confidence intervals, multiple seeds, and benchmark-stratified analysis.

Because LLM outputs are prompt-sensitive, all substantive claims should be made from repeated runs and aggregated distributions rather than single exemplars.

7 Execution and Reproducibility

The repository is already organized in a manner conducive to reproducible experimentation. It uses explicit experiment modules, provider abstraction for model access, configuration-driven prompt loading, automatic preflight testing, incremental result writing, and timestamped output artifacts that preserve both completed and interrupted runs. For the purposes of the proposal, the important point is not the exact command-line interface, but the fact that the experimental workflow is versioned, modular, and auditable.

Accordingly, the project should commit to version-controlled prompts and configurations, frozen prompt subsets for each run through stored metadata, retention of both pending and judged outputs for auditability, explicit separation of benchmarked models from judge models, reporting by provider, model, language, and benchmark source, and replicated runs for settings in which output variance is non-trivial. These commitments are more central to the scientific credibility of the project than a detailed inventory of engineering commands in the body of the proposal.

8 Why Hardware and API Credits Are Necessary

The request for hardware and API credits should be understood as a methodological requirement rather than as a matter of convenience.

8.1 Current Limitation

At present, the project can run meaningful pilot studies on the locally accessible open-weight cohort using a UIUC ICRN baseline of a single `1xH200` GPU. That baseline is valuable because it makes local experimentation possible, but it supports only pilot-scale runs, modest parameter sweeps, and a limited subset of the full experimental matrix. It is not sufficient to test the broader claim that altruistic or society-serving behavior differs between local open-weight systems and the contemporary frontier models most likely to be deployed in practice.

8.2 Why Frontier Access Matters

The core comparative question requires modern model families from OpenAI, Anthropic, and, after a modest provider integration, Google Gemini. These are closed-source API-served systems, and they cannot be evaluated at the scale required by this project without dedicated usage credits. Without them, the study would largely compare local open-weight models against one another. That is a useful starting point, but it does not answer whether the social behavior of the most capable commercial systems differs materially from the local baseline that the project can already access.

8.3 Why Credits Matter Even for Small-Text Experiments

Even though the individual prompts are short, the total call volume grows quickly. Part 0 currently implies tens of thousands of primary generations once the prompt set, language set, and model set are crossed, and the required judging pass substantially increases that total. Part 2 introduces a different but equally serious scaling problem: even a moderate society size and runtime produce thousands of agent decisions for a single condition, and the total rises rapidly once multiple models, multiple parameter settings, and multiple replications are introduced. When closed-source models from OpenAI, Anthropic, and Gemini are added to that matrix, API credits become a binding experimental requirement rather than a marginal convenience. Credits therefore determine not only whether the study can be run at all, but whether it can be run with adequate coverage, replication, and statistical credibility across both local and frontier cohorts.

8.4 Why Hardware Matters

Additional hardware is justified because it would allow the project to benchmark larger local open-weight baselines, accelerate multilingual batch evaluation, host local judge models when external API costs are undesirable, and execute the repeated generations, judge passes, parameter sweeps, and replications required by Parts 2 through 5 without impractical runtime. The current `1xH200` UIUC ICRN allocation is a strong starting point, but it is not enough to support the full local side of

the study at the same scale as the proposed frontier-model comparison. More compute is therefore needed for the open-weight arm of the project, just as API credits are needed for the closed-source arm. Hardware and credits are part of the experimental design itself. They are necessary to make the study contemporary, comparative, and statistically credible.

9 Ethics and Risk Management

This project necessarily uses harmful prompts because it studies robust refusal and misuse resistance, and that fact creates clear obligations. Harmful prompts should be stored and handled only for evaluation purposes; outputs should be reported in aggregate whenever possible; raw harmful completions should not be republished beyond what is necessary for secure internal analysis; benchmark usage should respect the norms and licenses of the upstream resources; and later phases involving simulated violence, theft, or exploitation should be framed explicitly as safety and governance evaluation rather than as behavioral endorsement. The proposal should also acknowledge an important conceptual limitation: "altruism" is used here as a behavioral construct, not as proof of moral agency, subjective intent, or genuine ethical understanding.

10 Deliverables

The project's deliverables should include:

1. a benchmark report for Part 0 with HarmBench and JBB breakdowns,
2. a game-theoretic cooperation report for Part 1,
3. a commons-sustainability report for Part 2,
4. an integrated cross-part analysis linking safety refusal to downstream social behavior,
5. a roadmap implementation plan for Parts 3–5,
6. plots, CSV outputs, metadata, and reproducible run configurations,
7. a final research paper or technical report suitable for academic submission.

11 Proposed Timeline

1. **Stage 1:** stabilize Part 0 reporting, benchmark splits, and multilingual analysis.
2. **Stage 2:** expand Part 1 and Part 2 replication across local safeguarded and uncensored models.
3. **Stage 3:** add frontier API comparison runs for OpenAI and Anthropic, then add Gemini support.
4. **Stage 4:** fit the cross-part statistical models and identify whether refusal predicts social cooperation.
5. **Stage 5:** implement reputation-aware Part 3, then begin the richer Part 4 and Part 5 society simulations.

12 Conclusion

This project is well-motivated because it links three questions that are often studied separately: whether a model refuses harmful requests, whether it cooperates under strategic tension, and whether a society of such models remains stable over time. The current repository already provides a substantive experimental backbone for Parts 0–2. The next step is to turn that foundation into a comparative, publishable study that includes local safeguarded models, local uncensored models, and frontier API models. Framing Part 0 explicitly around HarmBench and JailbreakBench, while extending Parts 1–5 as progressively richer social environments, gives the project a coherent scientific narrative and a strong basis for requesting the hardware and credits needed to execute it properly. Concretely, the study can already begin on the current UIUC ICRN `1xH200` baseline, but completing the intended comparison across both local open-weight and closed-source frontier systems requires additional compute and API credits.

References

[1] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. *Harm-*

Bench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. arXiv:2402.04249, 2024. https://arxiv.org/abs/2402.04249

[2] Center for AI Safety. *HarmBench GitHub Repository*. https://github.com/centerforaisafety/HarmBench

[3] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. *JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models*. arXiv:2404.01318, 2024. https://arxiv.org/abs/2404.01318

[4] JailbreakBench. *JailbreakBench GitHub Repository*. https://github.com/JailbreakBench/jailbreakbench

[5] JailbreakBench. *JailbreakBench Leaderboard and Behaviors*. https://jailbreakbench.github.io/behaviors

[6] Karthik Sreedhar, Alice Cai, Jenny Ma, Jeffrey V. Nickerson, and Lydia B. Chilton. *Simulating Cooperative Prosocial Behavior with Multi-Agent LLMs: Evidence and Mechanisms for AI Agents to Inform Policy Decisions*. arXiv:2502.12504, 2025. https://arxiv.org/abs/2502.12504

[7] Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Tao Ge, and Furu Wei. *ALYMPICS: LLM Agents Meet Game Theory – Exploring Strategic Decision-Making with AI Agents*. arXiv:2311.03220, 2023. https://arxiv.org/abs/2311.03220

[8] Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. *MultiAgentBench: Evaluating the Collaboration and Competition of LLM agents*. arXiv:2503.01935, 2025. https://arxiv.org/abs/2503.01935

[9] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. *Generative Agents: Interactive Simulacra of Human Behavior*. arXiv:2304.03442, 2023. https://arxiv.org/abs/2304.03442