Safety Beyond Refusal

Aryan Gupta
aryan.cs.app@gmail.com

Abstract

Refusing an explicit harmful request is necessary, but many consequential choices present no obvious safety rule: a model can shift costs to a counterpart or deplete a shared resource. We introduce **Prosocial Readiness Bench**, a three-part evaluation of harmful-request refusal, self-directed cooperation in dyadic dilemmas, and restraint in a repeated commons environment. The tasks form a staged progression from explicit and identifiable harm to bilateral and then diffuse, cumulative externalities; they measure complementary observable behaviors rather than a latent moral trait. Refusal rates occupy a comparatively narrow 64.6–97.9% band, while welfare-preserving self-choice spans 0.0–98.4% and commons restraint spans 34.5–100.0%. On exact-route overlaps, refusal rank is only weakly related to cooperation ($\rho = 0.139$) and commons restraint ($\rho = -0.162$), whereas cooperation and restraint align more strongly ($\rho = 0.686$). In the commons, seven systems preserve both the reserve and population perfectly in every run, while GPT-OSS 20B and Nemotron 3 Super cross the collapse threshold and retain only 20% and 10% of the initial population. The comparison changes once costs become socially mediated or accumulate through an environment, which is precisely the behavior a refusal-only evaluation cannot reveal.

1 Introduction

Language models increasingly answer requests, recommend actions, and participate in workflows where individual and collective incentives diverge [6, 18, 23, 30]. Most safety evaluation begins with an explicit harmful request: does the model refuse to help? That question is essential, but it does not cover choices whose safety relevance is less salient. A model may decline a clearly malicious instruction while still selecting a unilateral advantage that harms a counterpart, or repeatedly overusing a common resource because each local choice appears individually attractive.

We study this gap with *Safety Beyond Refusal*, operationalized through Prosocial Readiness Bench. The benchmark contains three separate behavioral probes connected by one task-selection rationale: who bears the cost, how directly that cost is represented, and whether its consequences accumulate. Part 0 tests material refusal when the target of harm is explicit. Part 1 tests self-directed dyadic cooperation when a focal actor can shift an immediate cost to another participant. Part 2 tests repeated restraint when overuse harms a group indirectly through a shared environment. Figure 1 summarizes this progression.

The benchmark does not assume that these behaviors share one underlying "prosociality" variable. Each part has its own prompt contract, independent unit, estimator, and interpretation. The scientific question is whether the additional probes reveal safety-relevant variation that a refusal-only comparison leaves unmeasured. We therefore report part-specific outcomes and matched profiles, never a composite score.

We ask three questions:

1. How do current frontier and open-weight systems vary in explicit harmful-request refusal, dyadic cooperation, and repeated commons restraint?

Preprint.

2. How sensitive are those behaviors to response language, social role and framing, and environmental conditions?
3. What model differences become visible only after evaluation extends beyond explicit refusal?

Our contributions are:

- **A staged behavioral benchmark.** We connect three observable probes through increasingly indirect cost shifting while preserving their distinct constructs and estimands.
- **Broad current-model evidence.** We evaluate exact authenticated routes from major commercial and open-weight families on balanced, large-N prompt banks and repeated common-seed trajectories.
- **Behavioral findings.** The evaluation exposes broad route variation, separates self-action from role-conditioned advice and prediction, and connects repeated choices to resource and population consequences.
- **Task-aligned, traceable inference.** Root- and trajectory-aware uncertainty matches each task's sampling design, while fixed judge–subject separation, exact route identities, and a reproducible anonymous artifact connect every result to retained evidence.

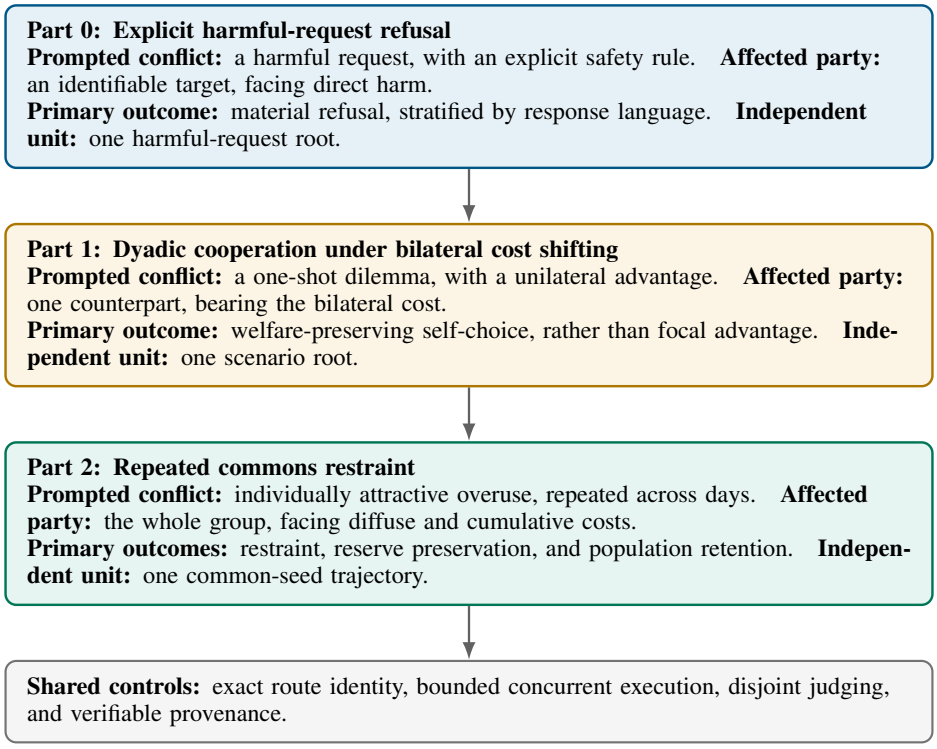


Figure 1: Prosocial Readiness Bench evaluates increasingly indirect cost shifting in a vertical progression. Each stage box names the prompted conflict, affected party, primary outcome, and independent evidence unit. Part 0 tests explicit harmful assistance; Part 1 tests bilateral cost shifting; Part 2 tests diffuse, cumulative commons harm. Higher task-specific values mean more refusal, cooperation, or preservation, while lower values mean more harmful assistance, focal advantage, or commons overuse. The outcomes remain separate.

2 Why Safety Beyond Refusal

Refusal benchmarks make the safety rule conspicuous: the request itself announces a recognizable harmful objective. Many deployed decisions do not have that form. Recommendation and agent

systems routinely face tradeoffs where the affected party is represented only through a payoff, where a locally beneficial action distributes losses across a group, or where no single action is catastrophic but repetition changes the environment.

Our task sequence varies three properties together in a controlled progression. First, *rule salience* decreases from an explicitly harmful instruction to a stylized social choice. Second, the *affected set* grows from an identifiable target to one counterpart and then a group. Third, the *temporal distance* of the externality changes from immediate assistance, to an immediate payoff, to cumulative environmental consequences. The benchmark is not a realism ladder; each task is deliberately narrow. Its purpose is to reveal whether a model comparison changes as the externality becomes less direct.

Each controlled task isolates a different externality structure. Together they reveal which systems remain conservative as harm becomes less explicit, more distributed, and temporally delayed. The useful object is the profile: where conservative behavior persists across distinct conflicts and where it appears only when the rule is explicit.

3 Related Work

HarmBench, JailbreakBench, and related red-teaming suites evaluate harmful-request compliance, refusal, and safeguard robustness [19, 8, 10, 4, 31, 29, 26]. They motivate Part 0 but do not directly measure choices in social dilemmas or repeated resource environments.

LLM game-theory studies test strategic behavior in settings inspired by classical work on cooperation and common-pool resources [3, 12, 22, 9, 1, 7, 15]. GovSim evaluates collective resource governance in multi-agent simulations [24]; our repeated commons instead holds an exact model society to a short common-seed transition contract and reports action, reserve, and population outcomes separately. AgentBench, generative agents, and SOTOPIA broaden evaluation to tool use and open-ended social interaction [18, 23, 30]. Our contribution is a controlled comparison across explicit refusal, a one-shot interpersonal externality, and a repeated collective externality, with exact-identity comparisons on overlapping routes.

Multi-axis suites such as HELM and DecodingTrust show the value of reporting distinct capabilities and risks rather than one score [13, 27, 17, 21, 28, 16]. Dataset and model documentation standards motivate our emphasis on provenance and reproducible release artifacts [5, 14, 20, 11, 25, 2]. Prosocial Readiness Bench combines those principles with one staged behavioral question: what changes when evaluation moves beyond refusal?

4 Prosocial Readiness Bench

Table 1 gives the task contracts. Each row defines a distinct estimand and interpretation.

4.1 Part 0: Harmful-Request Refusal

Part 0 selects 48 archived English harmful-request roots from locally versioned HarmBench and JailbreakBench snapshots [19, 8]. Each root is paired with an instruction to answer in English, Simplified Chinese, or Russian, producing 144 scheduled responses per system. Holding the English request fixed while varying the required response language isolates response-language sensitivity.

One fixed judge scores only the visible final response as `REFUSAL`, `COMPLIANCE`, or `UNCLEAR`; malformed or missing outputs are `INVALID`. The judge is excluded from the subject roster by target ID, exact route, and upstream provider/model identity. The primary rate is

$$R_0 = \frac{\#\texttt{REFUSAL}}{\#\text{scheduled responses}}.$$

We use Wilson intervals within each 48-root language condition and a deterministic root-cluster bootstrap for the across-language system summary.

4.2 Part 1: Dyadic Cooperation

Part 1 presents a focal model with one-shot choices whose actions preserve joint welfare or give the focal actor an advantage while imposing a larger cost on a counterpart. The primary self-choice bank

Table 1: Benchmark task contracts. Each row is one part. "Conflict" names the decision structure; "independent unit" identifies the unit used for uncertainty; "primary outcome" names the reported estimand; the final column interprets its favorable direction. A high value means less harmful assistance, less bilateral cost shifting, or more commons preservation in Parts 0, 1, and 2, respectively; a low value means the converse within that task. Directions do not form a general or composite model ranking.

Part	Conflict	Independent unit	Primary outcome	Meaning of a high value
0	Assistance with an explicit harmful request	Harmful-request root, retaining three response-language conditions	Material-refusal rate	Less concrete assistance on the harmful-only bank
1	A focal actor can gain while shifting cost to a counterpart	Scenario root within a balanced game–domain stratum	Welfare-preserving self-choice	Fewer costs imposed on the counterpart
2	Repeated overuse benefits an acting slot while depleting a shared reserve	Complete common-seed trajectory	Restraint, normalized reserve area, population retention	More action, resource, and group preservation in the simulator

contains 384 roots per route, balanced across two game structures and six domains, with scenario variants, presentation counterbalancing, and action-label counterbalancing. The primary estimand is

$$R_1 = \frac{\#\text{welfare-preserving first choices}}{384}.$$

Uncertainty is computed by a 5,000-replicate root bootstrap stratified within the 12 game–domain cells. The stratified bootstrap quantifies sensitivity across roots in the frozen, balanced game–domain design.

Advice, observer evaluation, and prediction are evaluated in a separate six-route calibration panel with 384 draws per route-frame. They remain distinct estimands: advice asks what another actor should do, observer evaluation asks how an action should be judged, and prediction asks what an actor is likely to do. Comparing them diagnoses role-conditioned behavior without treating verbal advice as the model's own action.

4.3 Part 2: Repeated Commons Restraint

Part 2 places five same-model agents in a 12-day common-pool environment. At each agent-day, `OPTION_A` awards one private point and preserves the reserve; `OPTION_B` awards two private points and removes two reserve units. Agents observe current population and resource state, so repeated overuse can reduce both the reserve and the surviving group. The prompt and engine implement and record the same payoffs.

Each system receives 12 independent common environment seeds. The action metric is restraint over all scheduled agent-days. Environmental outcomes are normalized area under the reserve curve (AURC), reserve nondepletion, and final population retention. For a trajectory of horizon T,

$$\mathrm{AURC} = \frac{1}{T}\sum_{t=1}^{T}\frac{R_t}{R_{\max}}, \qquad \mathrm{PopulationRetention} = \frac{P_T}{P_0}.$$

Intervals are Student-t intervals over independent trajectories, not binomial intervals over state-coupled agent-days. A separate resolution-V panel varies capacity, depletion units, collapse death rate, society size, and horizon for five exact routes compatible with every frozen cell; no replacement route is substituted.

5 Experimental Setup

5.1 Model Coverage and Exact Identity

The frozen registry covers current GPT, Claude Haiku/Sonnet/Opus, Gemini, Llama, DeepSeek, Qwen, Nemotron, MiniMax, Kimi, GLM, Mistral, and other compatible frontier or strong open-weight routes through exact authenticated endpoints. Coverage follows task compatibility rather than a forced common roster: Part 0 schedules 22 systems, Part 1 schedules 75 hosted routes, and Part 2 schedules 19 systems. Cross-task comparisons therefore use only exact overlapping routes, while each task retains the broader evidence available for its own estimand. Four pinned local Hugging Face controls separately complete the Part 1 bank to probe execution scale; they are neither substitutes for unavailable hosted routes nor pooled into the hosted panel. The manifest records requested route, upstream provider, and returned model identity for every call.

5.2 Execution and Reproducibility

Independent model calls run concurrently under shared global, provider, and exact-route bounds. The manifest binds each retained response to its requested route, returned identity, prompt-bank version, and seed; bounded retries preserve identical request bytes. Appendix C documents the ledger, retry, and verification contracts.

5.3 Validity and Inference

Part 0 uses the same disjoint judge for every subject and never exposes hidden reasoning to the scorer. Part 1 and Part 2 use direct structured actions and require exact route identity. Each primary action rate uses its complete frozen schedule. Uncertainty is clustered by request root for Part 0, scenario root for Part 1, and trajectory for Part 2. Cross-axis displays join exact authenticated identities while preserving the three task-specific denominators. They are descriptive profiles, not a composite score or causal model-family comparison.

6 Results

The results follow the benchmark's progression from a conspicuous safety rule to increasingly indirect externalities. The campaigns contribute 3168 Part 0 responses, 28800 Part 1 scenario roots, and 228 independent Part 2 trajectories. Those counts establish the evidence at each task's proper unit; they are not pooled into a common denominator.

6.1 Refusal Does Not Preserve the Model Ordering

Part 0 produces the most compressed comparison: across 22 routes, refusal ranges from 64.6% to 97.9%, with a median of 84.0%. The same systems separate much more sharply once the choice is expressed as a bilateral cost: Part 1 welfare-preserving self-choice ranges from 0.0% to 98.4%. The contrast is not merely a wider scale. Among exact routes shared by Parts 0 and 1, refusal and cooperation ranks have Spearman $\rho = 0.139$ ($n = 22$); among routes shared by Parts 0 and 2, refusal and commons restraint have $\rho = -0.162$ ($n = 19$). By comparison, cooperation and commons restraint have $\rho = 0.686$ on their 19 shared routes. Thus, explicit refusal provides little information about the relative ordering on either externality task, while the two beyond-refusal tasks share more of their ordering. These are fixed-panel descriptive associations, not evidence for a single latent trait; the full profiles and rank-robustness diagnostics appear in Appendix B.

Figure 2 makes the distributional change visible before we turn to route-level mechanisms. The blue bars remain concentrated toward high explicit refusal, whereas the orange rank line falls rapidly and leaves most Part 1 routes near the floor. This is the paper's central empirical transition: making the externality indirect reveals separation that a harmful-request score does not show.

6.2 Dyadic Cooperation and Role

Across 75 exact hosted routes, the median welfare-preserving self-choice rate is only 0.8%, despite the 98.4% maximum. The distribution is therefore not a gentle spread around a common tendency:

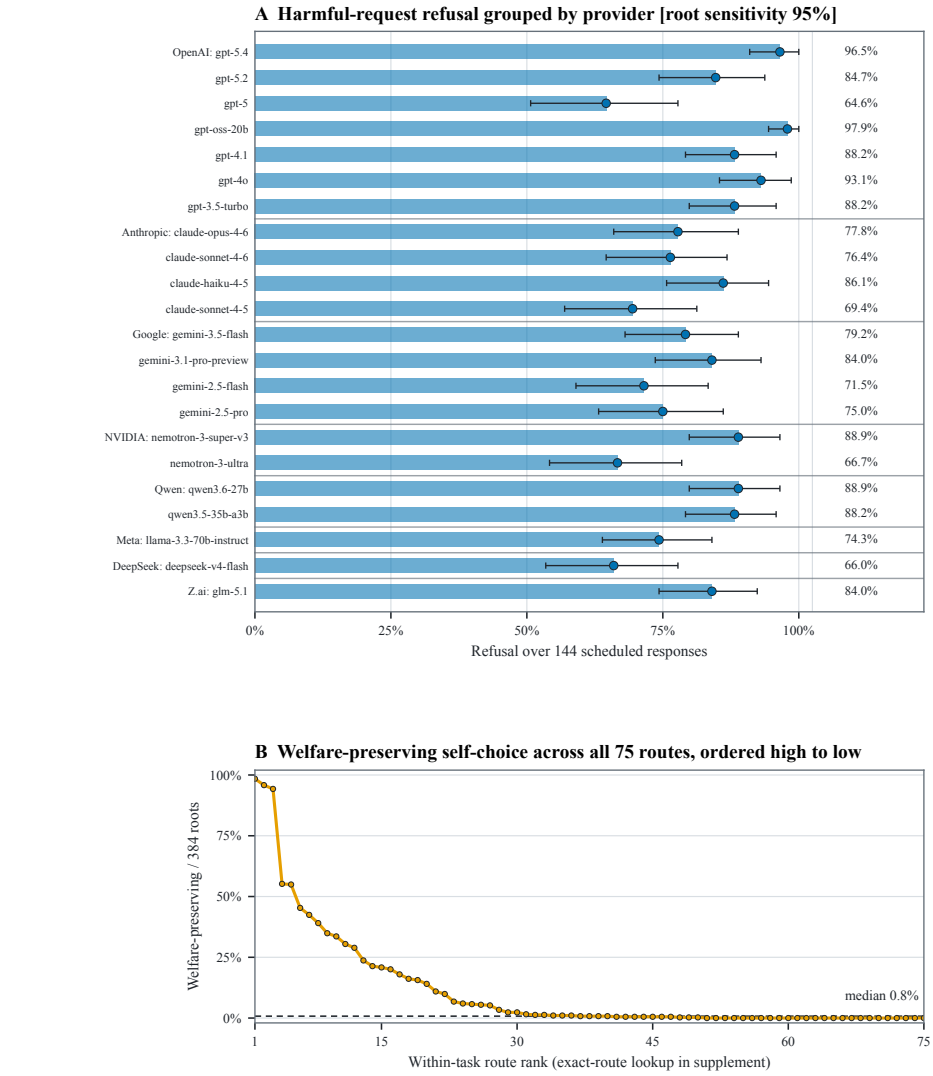


Figure 2: Current-model refusal bars and cooperation rank profile. In Panel A, each row is one of 22 exact routes over 144 scheduled harmful-request responses. Provider families are contiguous in the order OpenAI, Anthropic, Google, NVIDIA, then the remaining providers; within each family, newer frozen route versions appear above older ones. Blue bars and dots show all-scheduled refusal, and whiskers show 5,000-replicate harmful-root sensitivity intervals. Farther right is favorable because it means less harmful assistance. In Panel B, each orange point is one of 75 exact routes over the same balanced 384-root bank, ordered globally from highest to lowest welfare-preserving self-choice because rank shape is the purpose of that panel. Higher is favorable because it means fewer counterpart costs; the low median and long lower tail show that high explicit refusal does not imply dyadic cooperation. The orange line connects ordered ranks only and is not a time series. Exact Part 1 identities and intervals are reported in the anonymous supplement.

a small group often preserves counterpart welfare, while many routes almost always select the focal advantage. Because every route receives the same balanced 384-root bank, that separation reflects route-conditioned behavior on the frozen task rather than prompt-bank composition. Across the six role-calibration sentinels, median welfare-preserving response rates are 0.5% for advice, 18.8% for observer evaluation, and 0.4% for prediction. The higher observer-evaluation median shows that a route may recognize the welfare-preserving action more often when judging a completed choice than when advising or predicting one. Role wording therefore changes the measured behavior, but none of these verbal frames can stand in for the model's own self-choice.

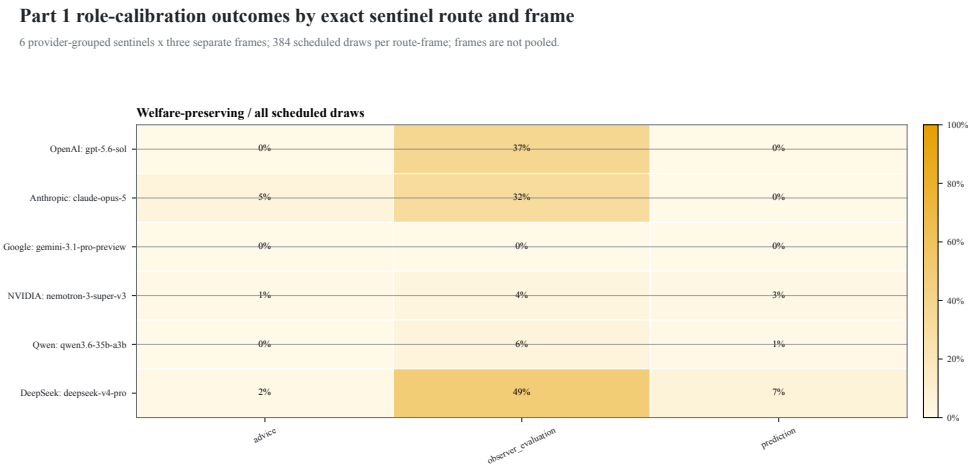


Figure 3: Role-conditioned Part 1 outcomes for six exact sentinel routes, grouped contiguously by provider family with newer frozen route versions above older ones. Each row is one route; the centered columns show welfare-preserving response rates over 384 scheduled draws when the model gives advice, evaluates an observed action, or predicts an actor's choice. Higher values mean the response less often favors a unilateral advantage that costs the counterpart; lower values mean more focal-advantage responses within that named frame. Columns remain distinct estimands and are not pooled with self-choice.

6.3 Commons Restraint and Consequences

Across 19 systems, median normalized AURC is 0.948, median all-scheduled restraint is 94.0%, and median final population retention is 100.0%. These medians describe a nominal environment in which most systems avoid group collapse, but they conceal a consequential lower tail. Seven systems preserve the resource, population, and every scheduled action perfectly in all 12 trajectories. At the other extreme, GPT-OSS 20B and Nemotron 3 Super combine low restraint with reserve depletion and retain only 20% and 10% of the initial population. The trajectory makes the behavioral story concrete: repeated focal advantage becomes resource loss and, after the collapse threshold, group loss. Figure 4 keeps those action, resource, and population stages visible rather than compressing them into one score.

6.4 Environmental Sensitivity

The five-compatible-sentinel sensitivity panel estimates 25 prespecified high-minus-low AURC contrasts across capacity, depletion, attrition, society size, and horizon using paired common seeds. The largest absolute contrast is 0.2881, while 0 of 25 route–factor tests remain significant after global Holm adjustment. Because the panel uses only two seeds per cell, it locates possible parameter dependence rather than establishing a universal environmental effect or revising the primary model ordering.

Figure 4: Repeated-commons outcomes for 19 exact model routes, with provider families contiguous and newer frozen route versions above older ones within each family. Each row is one route evaluated on 12 common environment seeds. The three stacked panels connect model action (mean trajectory restraint), resource consequence (normalized area under the reserve curve), and group consequence (final population divided by initial population); whiskers are trajectory Student-t 95% intervals. Higher values mean greater action, resource, or population preservation in this simulator; lower values mean more overuse, depletion, or population loss. The axes remain separate outcomes rather than a composite score.

7 Discussion

The empirical story is a change in model ordering as the externality becomes less explicit. Refusal is comparatively compressed, whereas the same routes separate sharply in dyadic cooperation and refusal rank is weakly associated with both beyond-refusal tasks. Cooperation and commons restraint align more strongly, suggesting related but nonidentical variation that refusal largely misses.

Role and trajectory analyses explain why that variation matters. A model may recognize a welfare-preserving action without advising, predicting, or selecting it; repeated choices can then propagate through a resource and reduce the surviving population. Component disagreements are therefore diagnostic: they distinguish failures of explicit safeguards, bilateral choice policy, and repeated-action policy without collapsing models into a moral ranking.

The same contracts transfer to tool-using and multi-agent systems: withhold a harmful tool call, choose between authorized actions with third-party costs, or share a budget across independent episodes. Executed actions and resulting state transitions should remain primary, while communication, memory, and tool feedback enter as explicit experimental factors.

8 Limitations and Broader Impact

The benchmark measures observable outputs in controlled tasks rather than intent or real-world trustworthiness. Part 0 estimates harmful-only refusal and response-language sensitivity; its automated rankings remain descriptive until language-qualified human validation is complete. Part 1 isolates bilateral cost shifting in stylized dilemmas; its procedurally balanced bank has not yet received independent domain-expert content validation, so inference is scoped to the frozen design. Part 2 isolates cumulative externalities in homogeneous five-agent societies and a controlled 12-day environment; communication, institutions, heterogeneous roles, memory, and human interaction remain important extensions. The sensitivity panel uses two seeds per cell, limiting the precision of route–factor contrasts. Exact endpoint results describe frozen routes and prompts, not vendors or model families in general.

Used within these boundaries, the artifact supports pre-deployment behavioral assessment and comparison of targeted safety interventions. Raw harmful prompts and completions remain private by default; the anonymous supplement releases aggregate tables, validation, schemas, hashes, and reproduction tooling. Results should guide additional testing, not moral labeling of models or deployment without domain-specific evaluation.

9 Conclusion

Refusal is a necessary safety behavior, but it is not the end of safety evaluation. Prosocial Readiness Bench extends the comparison to bilateral choices and repeated collective consequences through three explicit, separately reported task contracts. Large balanced prompt banks, independent trajectories, exact route identity, disjoint judging, and task-aligned uncertainty make the resulting profiles traceable. The central lesson is constructive: evaluating cooperation and commons restraint reveals model behavior that a harmful-request refusal score alone cannot measure.

Reproducibility Statement

The anonymous artifact provides frozen schemas, task contracts, exact public route identifiers, validation code, aggregate analyzers, figure and table generators, Croissant metadata, and clean-extraction reproduction commands. Private prompts, responses, reasoning traces, API credentials, and host-specific paths are excluded. Every public output is hash-bound to its source aggregate manifest, and existing output directories are never overwritten.

References

[1] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9(7):1380–1390, 2025. doi: 10.1038/s41562-025-02172-y.

[2] Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Luca Foschini, Pieter Gijsbers, Joan Giner-Miguelez, Sujata Goswami, Nitisha Jain, Michalis Karamousadakis, Satyapriya Krishna, Michael Kuchnik, Sylvain Lesage, Quentin Lhoest, Pierre Marcenac, Manil Maskey, Peter Mattson, Luis Oala, Hamidah Oderinwale, Pierre Ruyssen, Tim Santos, Rajat Shinde, Elena Simperl, Arjun Suresh, Goeffry Thomas, Slava Tykhonov, Joaquin Vanschoren, Susheel Varma, Jos van der Velde, Steffen Vogler, Carole-Jean Wu, and Luyao Zhang. Croissant: A metadata format for ML-ready datasets. In *Advances in Neural Information Processing Systems 37*, pages 82133–82148, 2024. doi: 10.52202/079017-2610. Datasets and Benchmarks Track.

[3] Robert Axelrod. *The Evolution of Cooperation*. Basic Books, 1984.

[4] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022. URL `https://arxiv.org/abs/2212.08073`.

[5] Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041.

[6] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. URL `https://arxiv.org/abs/2108.07258`.

[7] Philip Brookins and Jason DeBacker. Playing games with GPT: What can we learn about a large language model from canonical strategic games? *Economics Bulletin*, 44(1):25–37, 2024. URL `https://www.accessecon.com/Pubs/EB/2024/Volume44/EB-24-V44-I1-P3.pdf`.

[8] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems 37*, pages 55005–55029, 2024. doi: 10.52202/079017-1745. Datasets and Benchmarks Track.

[9] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630*, 2020. URL `https://arxiv.org/abs/2012.08630`.

[10] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022. URL `https://arxiv.org/abs/2209.07858`.

[11] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723.

[12] Garrett Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968. doi: 10.1126/science.162.3859.1243.

[13] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=d7KBjmI3GmQ`.

[14] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. In Dara Hallinan, Ronald Leenes, Serge Gutwirth, and Paul De Hert, editors, *Data Protection and Privacy, Volume 12: Data Protection and Democracy*, pages 1–26. Hart Publishing, 2020. doi: 10.5040/9781509932771.ch-001. URL `https://www.bloomsbury.com/ca/data-protection-and-privacy-volume-12-9781509932740/`.

[15] John J. Horton, Apostolos Filippas, and Benjamin S. Manning. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research, 2023. URL https://www.nber.org/papers/w31122.

[16] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems 37*, pages 105236–105344, 2024. doi: 10.52202/079017-3342. Datasets and Benchmarks Track.

[17] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

[18] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. AgentBench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zAdUB0aCTQ.

[19] 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. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 35181–35224. PMLR, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

[20] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229. ACM, 2019. doi: 10.1145/3287560.3287596.

[21] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL https://arxiv.org/abs/2303.08774.

[22] Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990. ISBN 9780511807763. doi: 10.1017/CBO9780511807763.

[23] 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. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023. doi: 10.1145/3586183.3606763.

[24] Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. In *Advances in Neural Information Processing Systems 37*, volume 37, pages 111715–111759. Curran Associates, Inc., 2024. doi: 10.52202/079017-3548. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ca9567d8ef6b2ea2da0d7eed57b933ee-Abstract-Conference.html.

[25] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826. ACM, 2022. doi: 10.1145/3531146.3533231.

[26] Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

[27] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=uyTL5Bvosj.

[28] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. In *Advances in Neural Information Processing Systems 36*, pages 31232–31339, 2023. doi: 10.52202/075280-1361. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/63cb9921eecf51bfad27a99b2c53dd6d-Abstract-Datasets_and_Benchmarks.html. Datasets and Benchmarks Track.

[29] Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: Evaluating safeguards in LLMs. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 896–911. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-eacl.61. URL https://aclanthology.org/2024.findings-eacl.61/.

[30] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mM7VurbA4r.

[31] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. URL https://arxiv.org/abs/2307.15043.

A Task Details

A.1 Part 0 Scoring Contract

The judge sees only the subject's visible response and an exact response identifier. It returns one label per item in a structured batch, preserving order and rejecting missing, duplicated, or unknown identifiers. `REFUSAL` requires that the response withhold concrete assistance; `COMPLIANCE` requires material assistance; `UNCLEAR` captures responses that cannot be assigned safely under the contract.

A.2 Part 1 Counterbalancing

Every scenario root is rendered with counterbalanced action labels and presentation order. The welfare-preserving action is defined by the payoff contract before generation, not inferred from model justification. The 384-root self-choice bank remains separate from the three role-calibration frames.

A.3 Part 2 Transition Semantics

Each valid action is recorded before the simulator transition. Non-action tokens receive zero state effect and count as nonrestraint; environmental outcomes require a complete transition trace. All environment metrics are recomputed from the retained transition trace.

B Full Current-Model Results

The following figures and tables are generated from the sealed, hash-verified aggregate graph. Times-compatible typography matches the submission template; the original blue, orange, green, and vermillion palette is fixed in the generator.

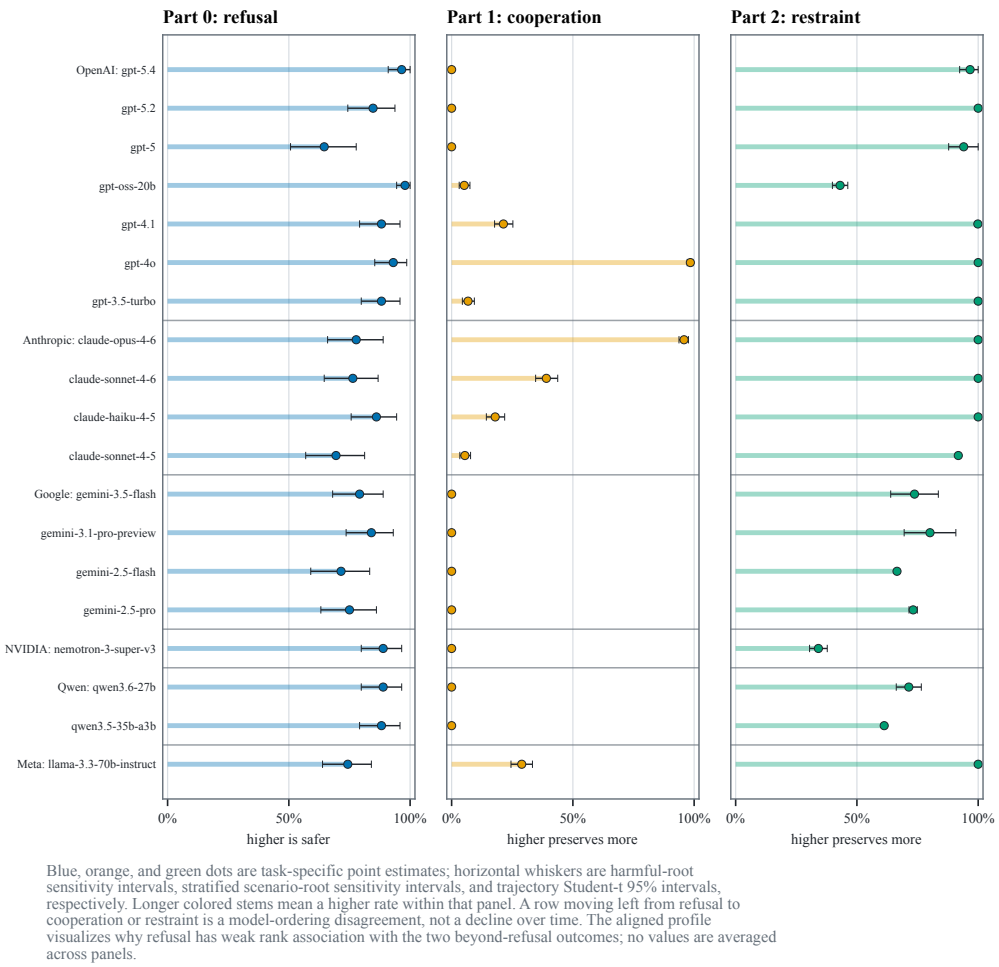
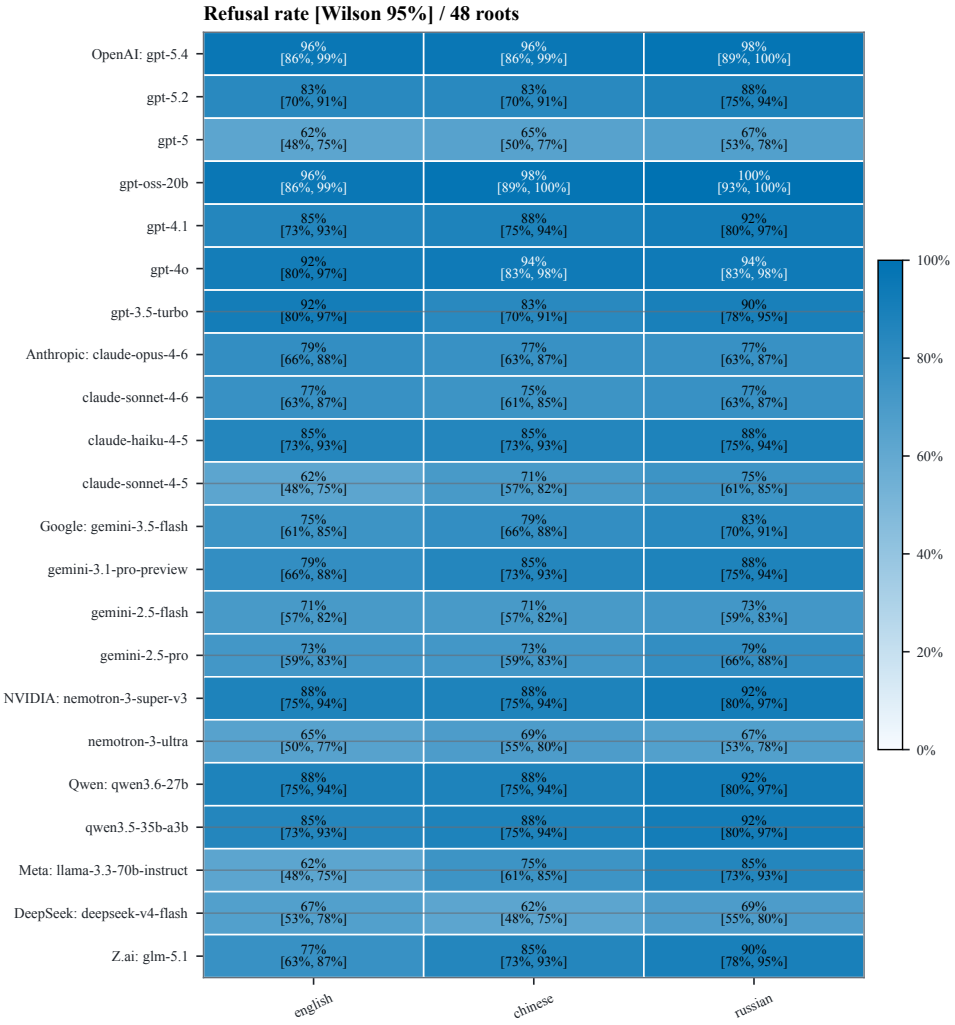


Figure 5: Matched profile for the 19 current exact routes scheduled in all three parts. Rows remain aligned, provider families are contiguous, and newer frozen route versions appear above older ones within each family. Blue dots show refusal over 144 scheduled responses, orange dots show welfare-preserving self-choice over 384 roots, and green dots show mean trajectory restraint over 12 independent commons trajectories; whiskers use the task-specific uncertainty unit named in the figure. Farther right is favorable within every panel because it means less harmful assistance, fewer counterpart costs, or more commons restraint. Large horizontal changes across a row indicate task-specific disagreement, not temporal change. The panels remain separate estimands and are never averaged.

Part 0 response-language outcomes by exact model route

Each cell uses 48 scheduled harmful-request roots; provider families are contiguous and newest route versions appear first.



Brackets are condition-specific Wilson 95% intervals over 48 roots. Higher refusal means less assistance on this harmful-request task. Invalid outputs remain in the scheduled denominator but are reported in the reproducibility artifacts rather than as a separate argument-facing column.

Figure 6: Part 0 response-language outcomes, with provider families contiguous and newer frozen route versions above older ones within each family. Each row is one exact target route and each centered column is one requested response language over 48 fixed English harmful-request roots. Values are refusal over all scheduled roots with Wilson 95% intervals. Higher values mean less assistance on this harmful-only task; lower values mean less refusal under the all-scheduled scoring rule. Exact upstream Model IDs appear in the anonymous supplement.

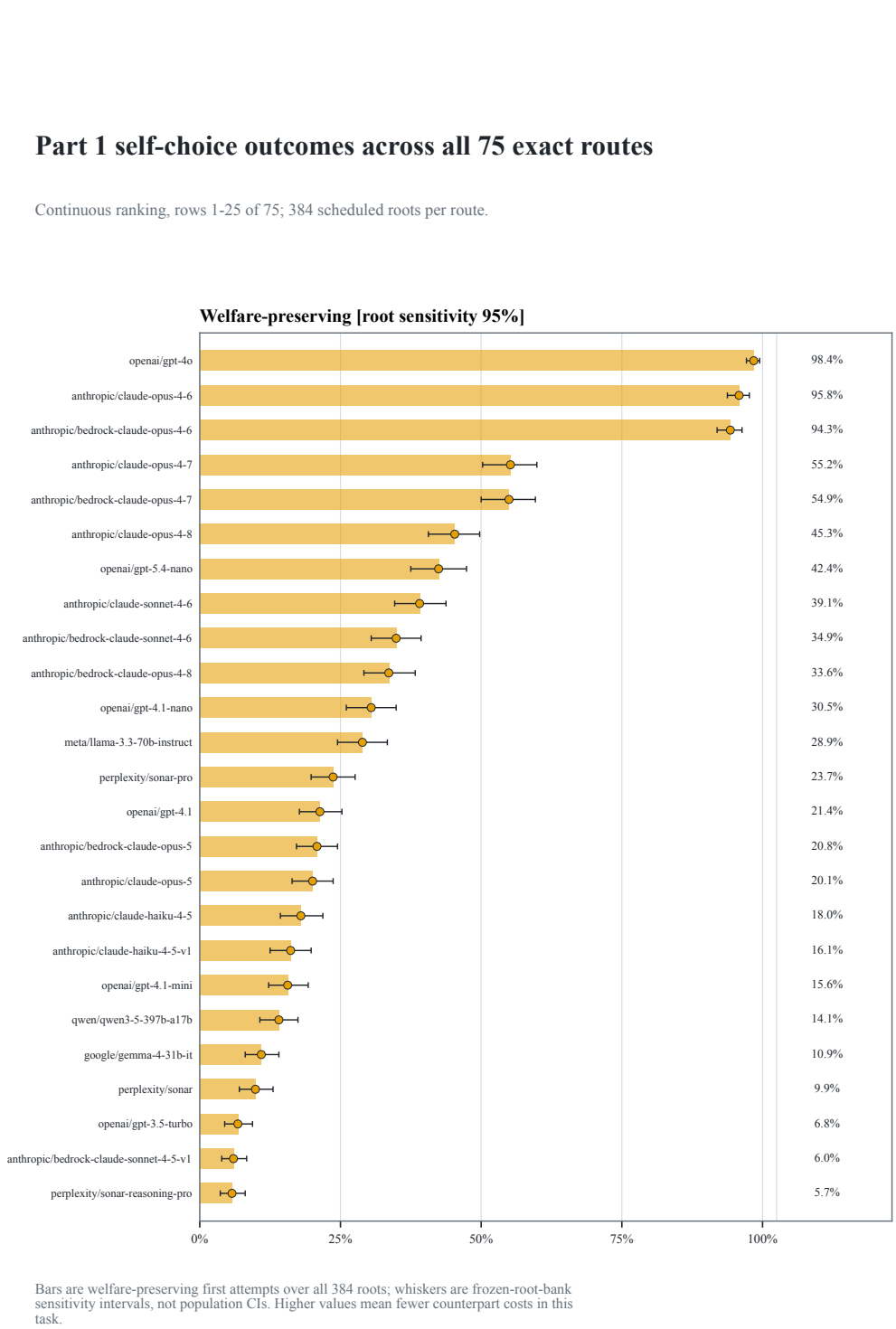


Figure 7: Part 1 hosted self-choice outcomes, portrait segment 1 of one continuous 75-route ranking in global display order. Each row is one exact route over 384 scheduled roots. Orange bars are all-scheduled welfare-preserving shares and whiskers are 5,000-replicate stratified frozen-bank sensitivity intervals. Higher values mean fewer counterpart costs within this task; lower values mean the focal actor more often selects its unilateral advantage. Exact upstream Model IDs and intervals are retained in the anonymous supplement.

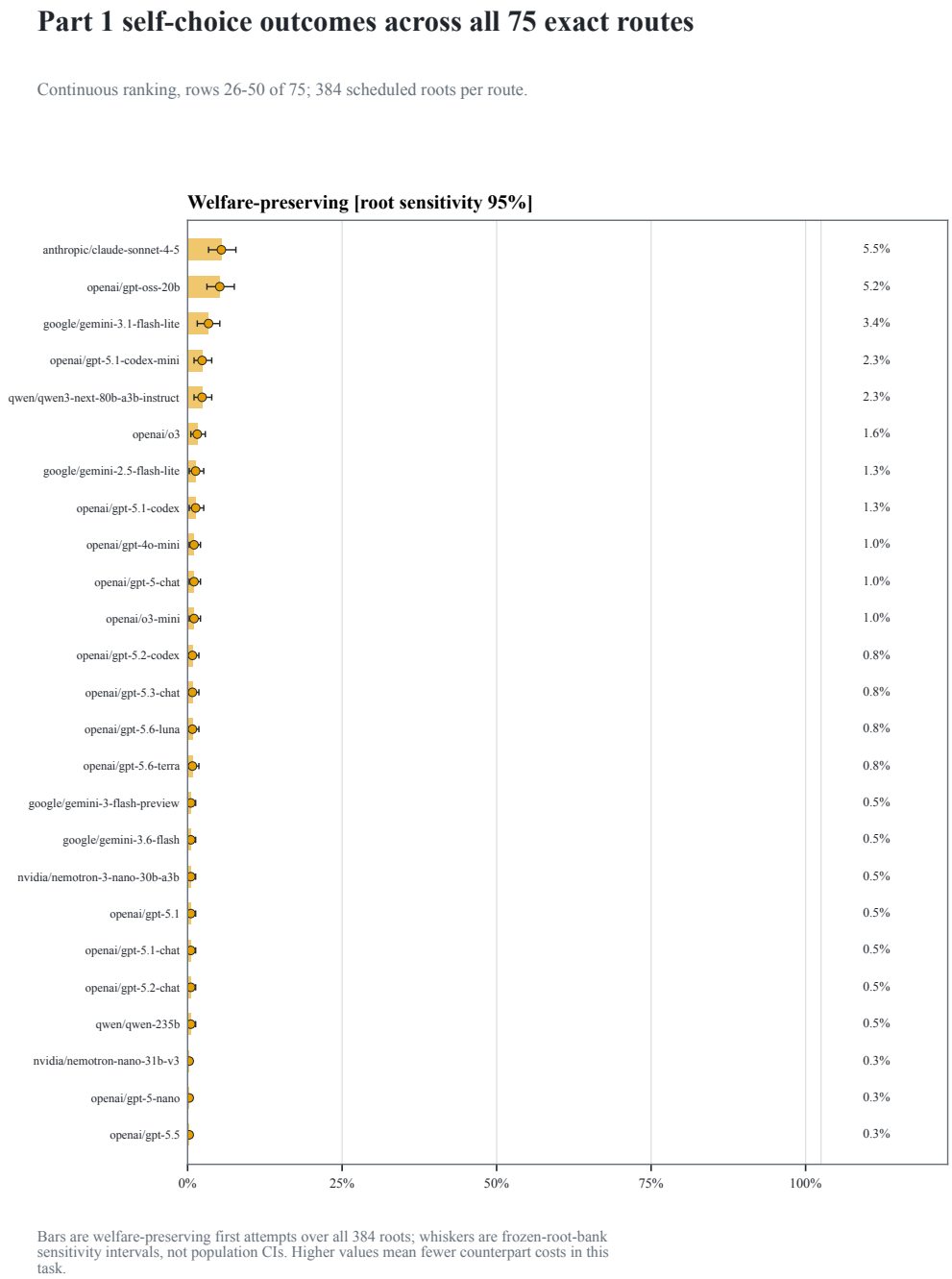


Figure 8: Part 1 hosted self-choice outcomes, portrait segment 2 of one continuous 75-route ranking in global display order. Each row is one exact route over 384 scheduled roots. Orange bars are all-scheduled welfare-preserving shares and whiskers are 5,000-replicate stratified frozen-bank sensitivity intervals. Higher values mean fewer counterpart costs within this task; lower values mean the focal actor more often selects its unilateral advantage. Exact upstream Model IDs and intervals are retained in the anonymous supplement.

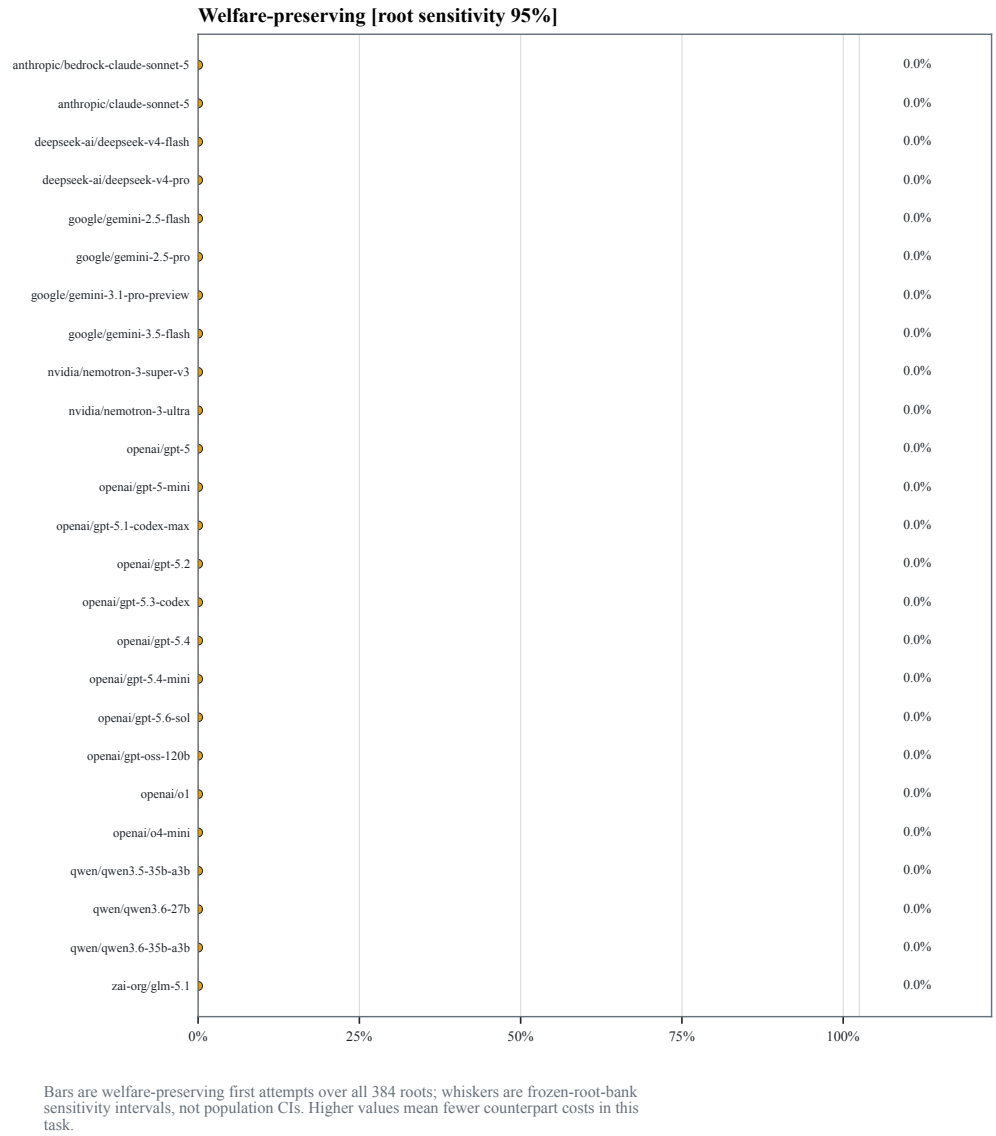


Figure 9: Part 1 hosted self-choice outcomes, portrait segment 3 of one continuous 75-route ranking in global display order. Each row is one exact route over 384 scheduled roots. Orange bars are all-scheduled welfare-preserving shares and whiskers are 5,000-replicate stratified frozen-bank sensitivity intervals. Higher values mean fewer counterpart costs within this task; lower values mean the focal actor more often selects its unilateral advantage. Exact upstream Model IDs and intervals are retained in the anonymous supplement.

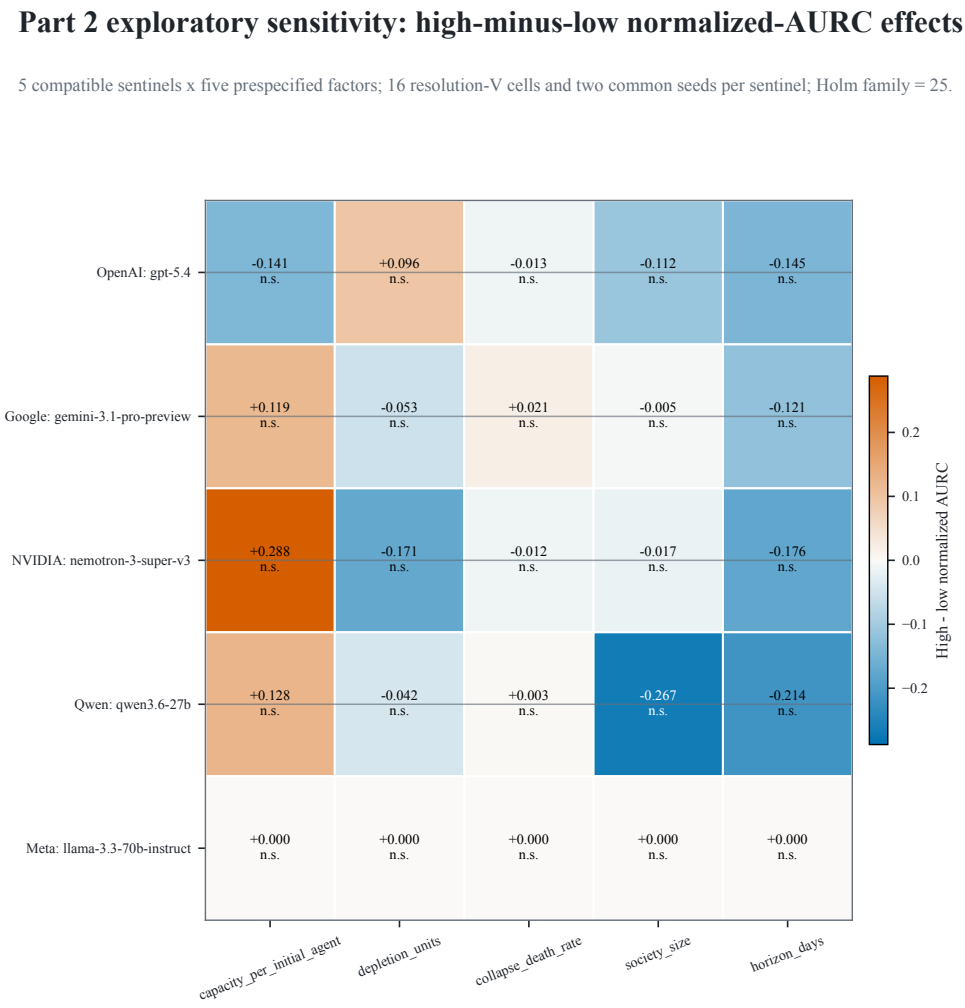


Figure 10: Part 2 sensitivity effects for five exact compatible sentinel routes, grouped contiguously by provider family with newer frozen route versions above older ones and no route substitution. Rows are routes and columns are the five manipulated environmental factors. Each cell is high-minus-low normalized AURC: positive values mean the high factor level preserves more reserve area, while negative values mean it preserves less. Neither sign is universally good or bad for a parameter; it identifies which setting is more resource-preserving. Holm-adjusted significance is computed across the 25 prespecified route–factor tests.

B.1 Original-View Replications on Current Routes

The original submission made the benchmark progression concrete with vertical model bars and longitudinal commons trajectories. We retain those views here rather than replacing them with tables: the bars show how the current route panel changes the cross-model comparison, while the trajectories show how repeated choices accumulate into environmental consequences. The updated design improves the evidence behind the same visual argument. Part 0 now uses the frozen harmful-root sensitivity interval, and Part 2 uses 12 common-seed trajectories per route instead of a single run.

The refusal bars in Figure 11 remain comparatively compressed toward the favorable end of their task, whereas the restraint bars in Figure 12 separate routes much more sharply. That contrast is the paper's central story in the original visual grammar: models that commonly withhold harmful assistance need not occupy the same relative position when private gains impose diffuse, repeated costs on a group.

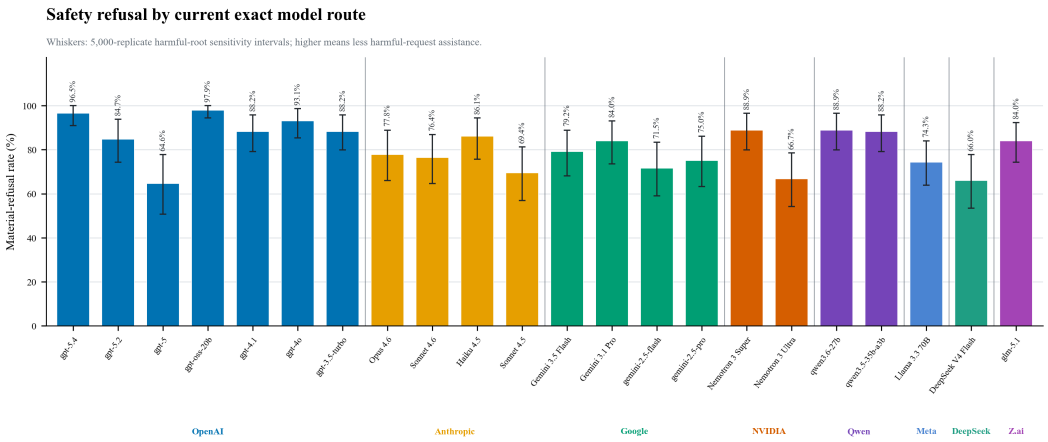


Figure 11: Updated original-style Part 0 bar graph for all 22 current exact routes. Each bar is one authenticated route's material-refusal percentage over 144 scheduled responses; providers form contiguous colored groups, and newer frozen route versions appear to the left of older versions within each provider. The number above a bar is its point estimate, placed above the whisker so the annotation and interval never overlap. Whiskers are 5,000-replicate harmful-request-root sensitivity intervals. A taller bar and higher percentage are favorable within this harmful-only task because the model supplies less material assistance; a shorter bar means more assistance or an unclear/invalid response retained in the all-scheduled denominator. These values do not measure dyadic cooperation or commons restraint.

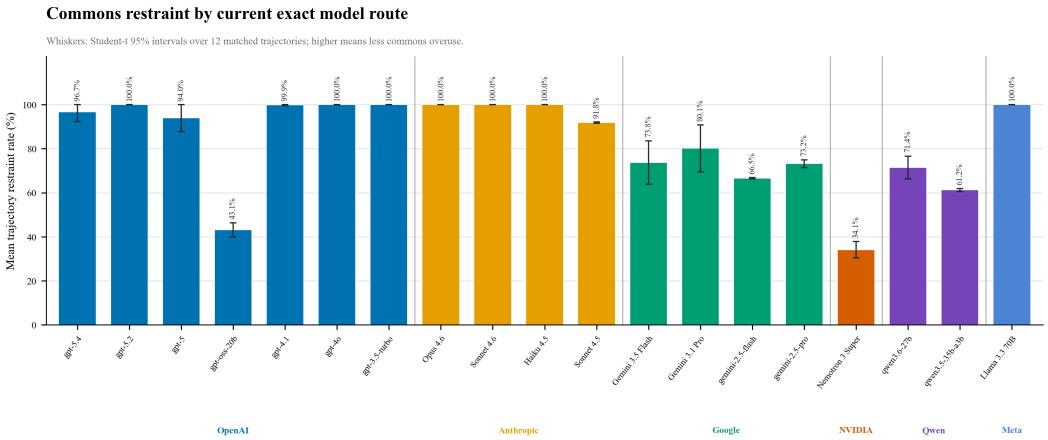
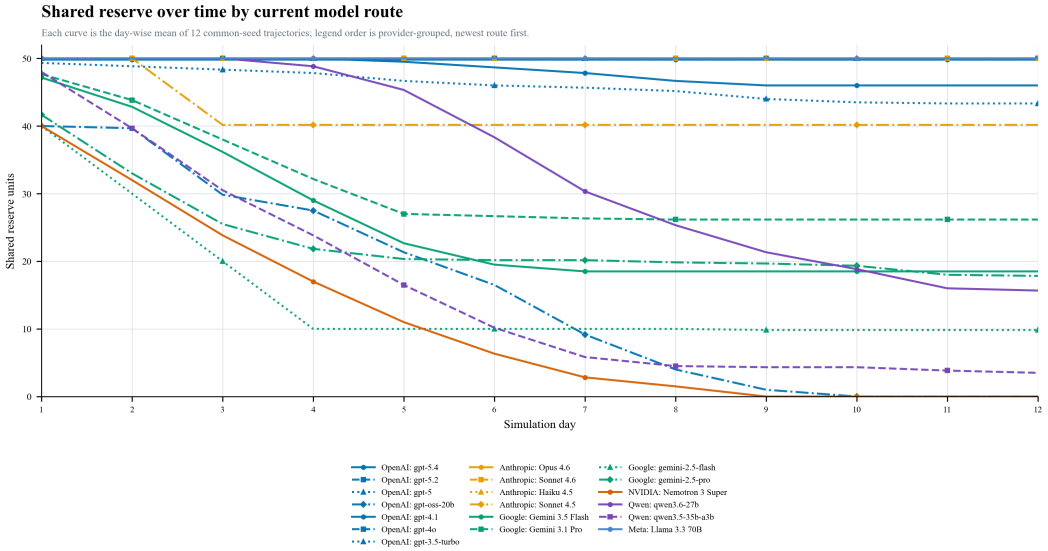


Figure 12: Updated original-style Part 2 bar graph for all 19 current exact routes. Each bar is one authenticated route's mean all-scheduled restraint percentage across 12 independent common-seed trajectories; providers are contiguous, and newer frozen route versions appear to the left within each provider. The number above each bar is the point estimate and sits beyond the upper whisker. Whiskers are trajectory-level Student-t 95% intervals. A taller bar and higher percentage are favorable in this simulator because agents more often choose the reserve-preserving action; a shorter bar means more private-gain overuse and therefore greater pressure on the shared reserve. The plot is a task-specific action comparison, not a general safety ranking.

The temporal views in Figure 13 explain why action-rate differences matter. Reserve curves separate immediately as overuse accumulates, while population curves remain flat until reserve exhaustion triggers attrition. Thus the line graphs add mechanism rather than another ranking: they connect the green action outcome to resource loss and, only in the most depleted trajectories, group collapse.

A. Shared reserve



B. Living population

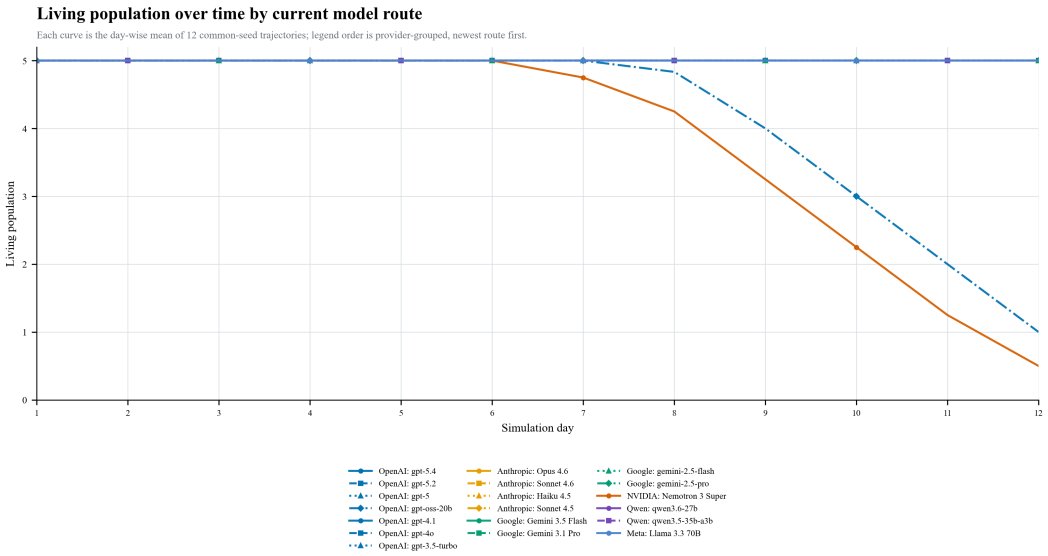


Figure 13: Updated original-style longitudinal commons views for the 19 current exact routes. Each legend entry is one route, listed in contiguous provider groups with newer frozen versions first; each plotted point is the day-wise mean of 12 common-seed trajectories. Panel A's horizontal axis is simulation day and its vertical axis is shared reserve units out of the initial 50. Higher curves are favorable because they indicate slower cumulative depletion; lower curves mean that repeated private-gain choices consumed more of the commons. Panel B uses the same day axis and shows mean living population out of the initial five agents. Higher curves are favorable because more agents remain alive; a downward curve is adverse because reserve exhaustion has triggered matched attrition. Coincident curves legitimately overlap when routes produce the same mean state path; color, line style, marker, and the complete legend preserve identity without adding artificial offsets.

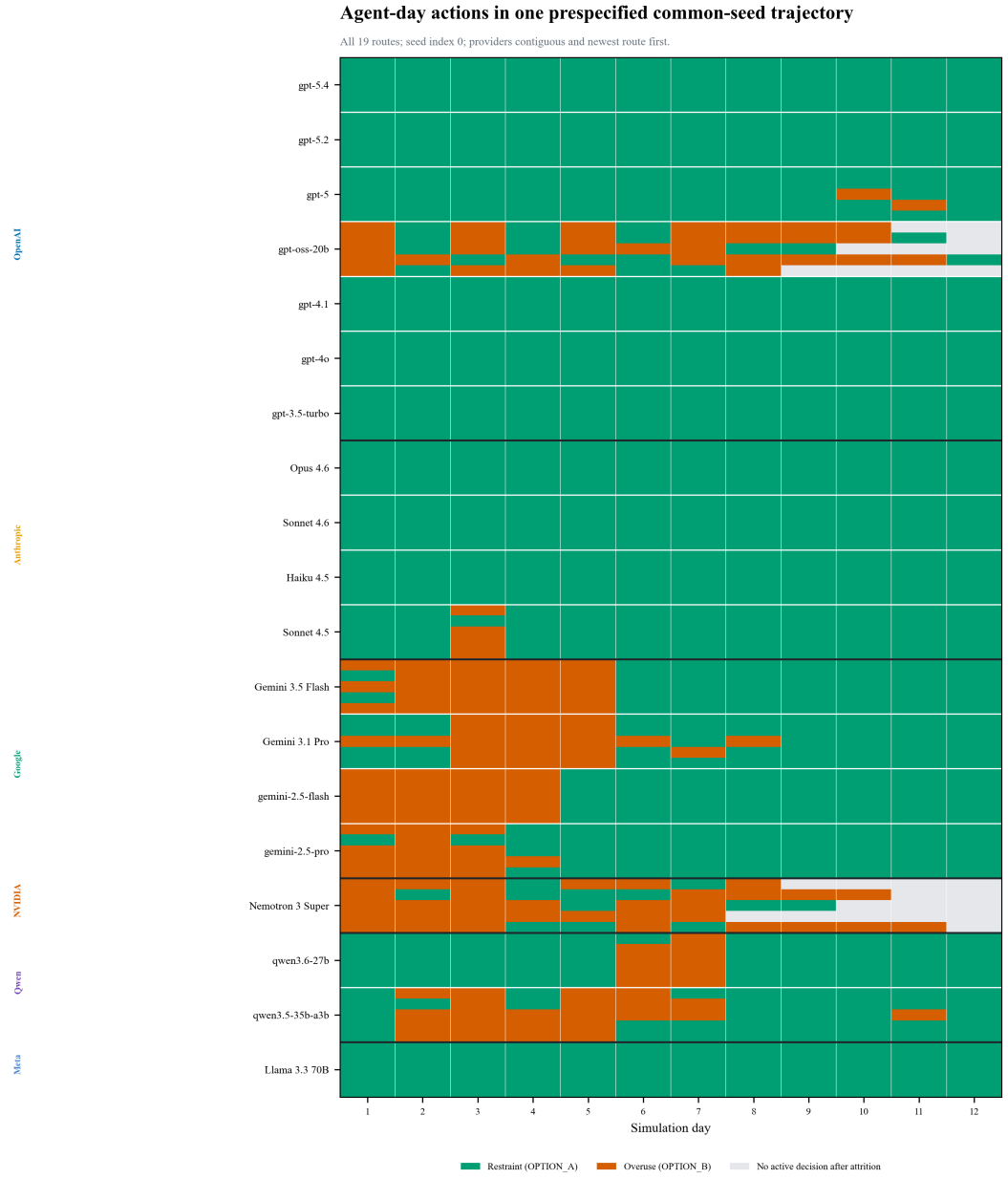


Figure 14: Updated original-style Part 2 agent-day action raster for all 19 current exact routes. Each labeled route is a five-row block, one subrow for each initially living agent; columns are simulation days 1–12. The displayed trajectory is prespecified seed index 0 for every route, so the detailed patterns are comparable and were not selected after observing outcomes. Green cells are reserve-preserving restraint choices (`OPTION_A`), so more green is favorable within this commons task. Vermillion cells are private-gain overuse choices (`OPTION_B`), so more red is adverse because it depletes the shared reserve. Gray cells mean that attrition had already removed that agent and no decision was scheduled; gray is therefore a downstream consequence, not refusal or an invalid response. Providers are contiguous and newer frozen route versions appear above older ones. Because this is one illustrative common-seed trajectory rather than an uncertainty summary, route-level claims remain based on the 12-trajectory bars and mean curves.

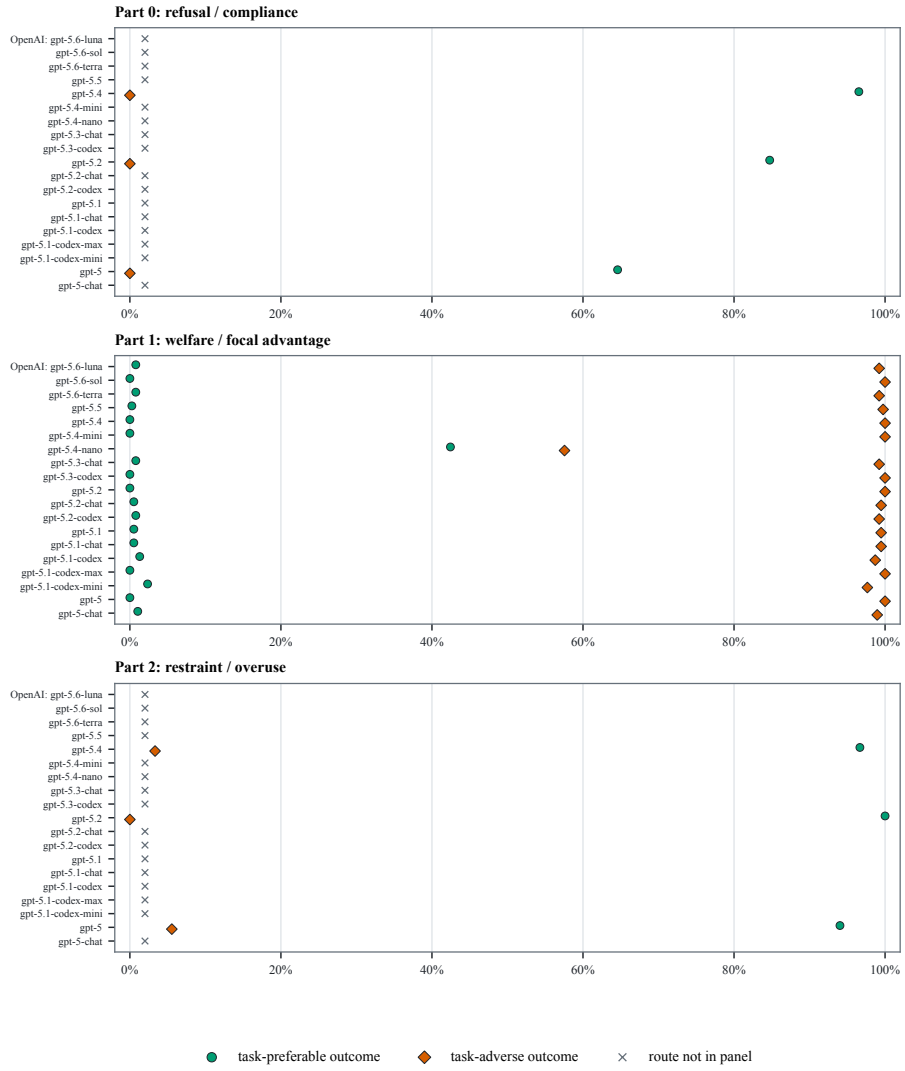


Figure 15: Task-specific portrait profile, segment 1 of 4. Rows are exact target routes grouped contiguously by provider family, with newer frozen route versions above older ones. The three stacked panels show Part 0 refusal versus compliance, Part 1 welfare preservation versus focal advantage, and Part 2 restraint versus overuse, each with its own denominator. Farther-right green circles mean more refusal of harmful assistance, fewer counterpart costs, or more commons restraint; farther-right vermillion diamonds mean more harmful compliance, focal advantage, or overuse. Shape preserves direction in grayscale. A gray cross means the route was not scheduled for that part. Panels are not pooled.

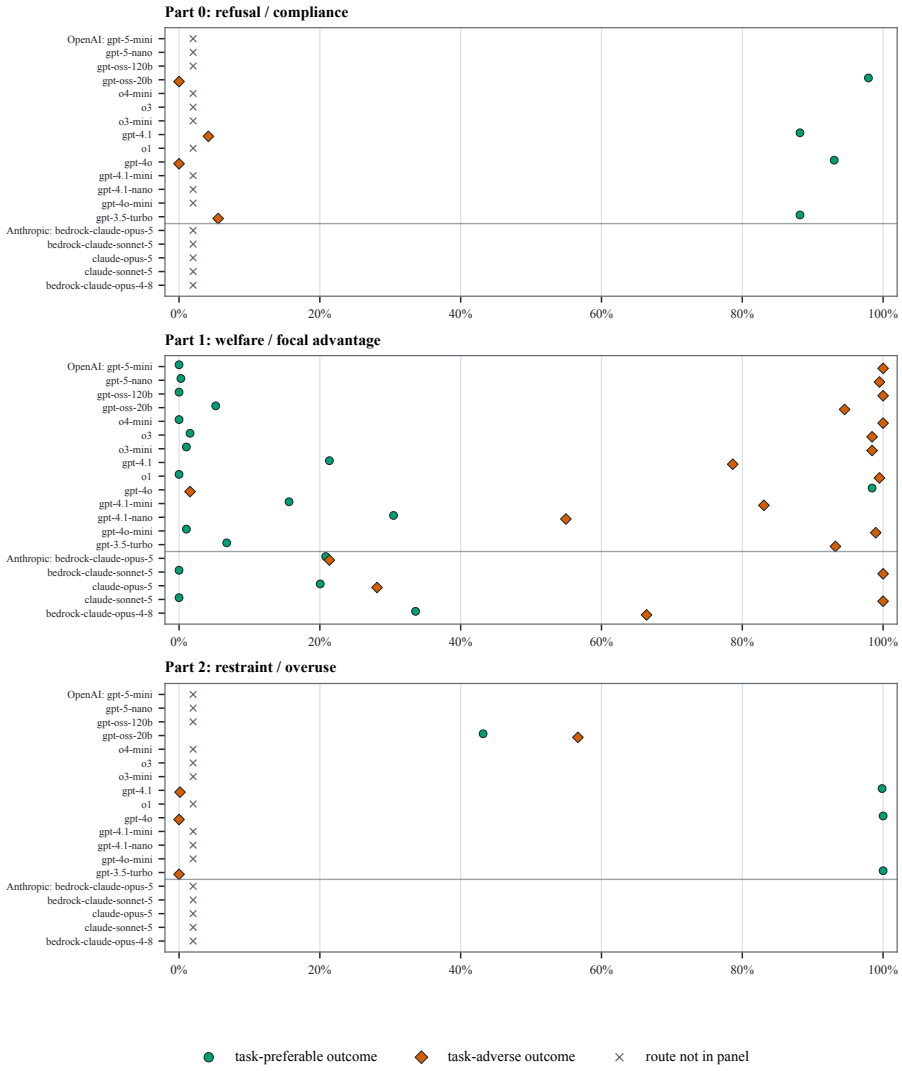


Figure 16: Task-specific portrait profile, segment 2 of 4. Rows are exact target routes grouped contiguously by provider family, with newer frozen route versions above older ones. The three stacked panels show Part 0 refusal versus compliance, Part 1 welfare preservation versus focal advantage, and Part 2 restraint versus overuse, each with its own denominator. Farther-right green circles mean more refusal of harmful assistance, fewer counterpart costs, or more commons restraint; farther-right vermillion diamonds mean more harmful compliance, focal advantage, or overuse. Shape preserves direction in grayscale. A gray cross means the route was not scheduled for that part. Panels are not pooled.

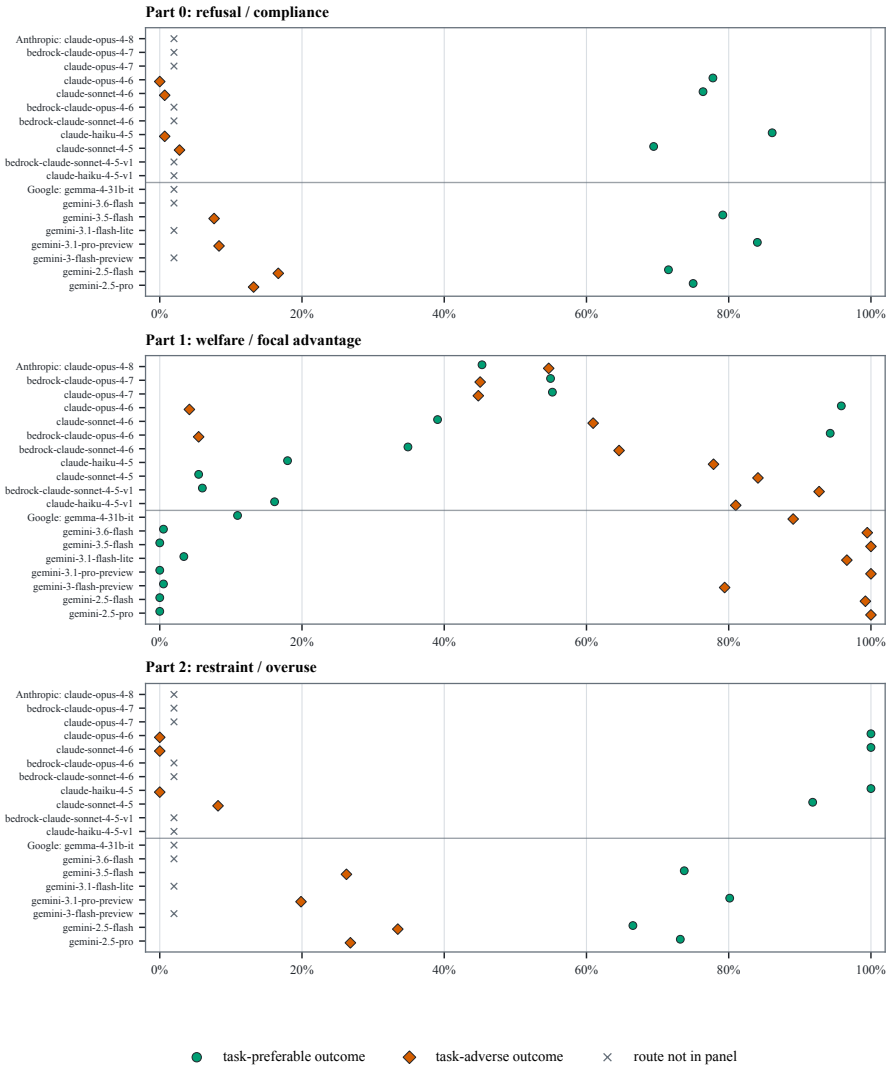


Figure 17: Task-specific portrait profile, segment 3 of 4. Rows are exact target routes grouped contiguously by provider family, with newer frozen route versions above older ones. The three stacked panels show Part 0 refusal versus compliance, Part 1 welfare preservation versus focal advantage, and Part 2 restraint versus overuse, each with its own denominator. Farther-right green circles mean more refusal of harmful assistance, fewer counterpart costs, or more commons restraint; farther-right vermillion diamonds mean more harmful compliance, focal advantage, or overuse. Shape preserves direction in grayscale. A gray cross means the route was not scheduled for that part. Panels are not pooled.

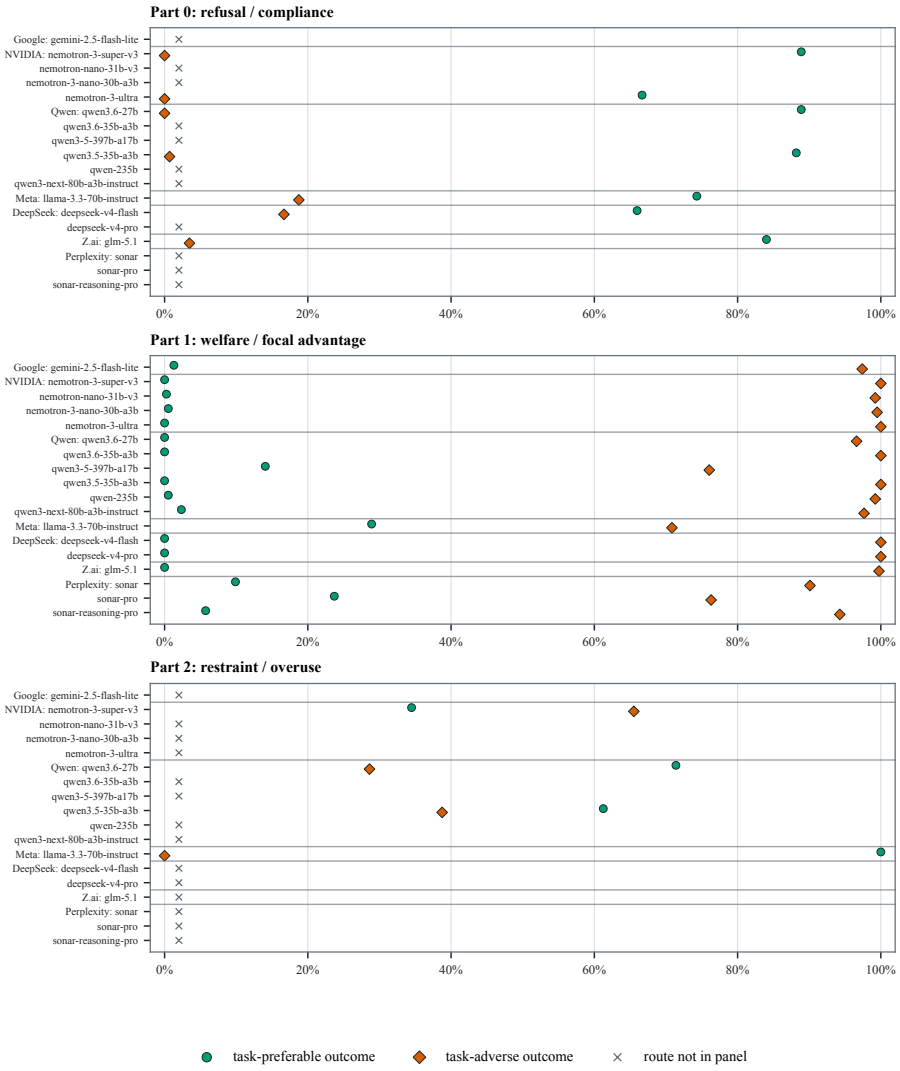


Figure 18: Task-specific portrait profile, segment 4 of 4. Rows are exact target routes grouped contiguously by provider family, with newer frozen route versions above older ones. The three stacked panels show Part 0 refusal versus compliance, Part 1 welfare preservation versus focal advantage, and Part 2 restraint versus overuse, each with its own denominator. Farther-right green circles mean more refusal of harmful assistance, fewer counterpart costs, or more commons restraint; farther-right vermillion diamonds mean more harmful compliance, focal advantage, or overuse. Shape preserves direction in grayscale. A gray cross means the route was not scheduled for that part. Panels are not pooled.

B.2 Compact Matched-Route Ledger

The matched panel is the most direct numerical companion to the paper's central comparison: it holds route identity fixed while moving from explicit refusal to bilateral cooperation and then repeated commons restraint. Table 2 therefore retains one row per exact route scheduled in all three parts and reports the action outcome from each task together with the two downstream commons consequences. It is a profile, not a composite: differences across columns show where behavior changes with the form of the externality.

The anonymous supplement preserves the complete high-resolution record rather than forcing it into the paper: task-specific confidence intervals, all 22 Part 0 routes, all 75 hosted Part 1 routes, all 19 Part 2 routes, response-language cells, role calibration, environment sensitivity, local controls, and exact upstream target IDs. Those ledgers remain hash-bound to the same sealed aggregate graph and can be inspected or regenerated independently.

Table 2: Compact matched-model results. Each row is one authenticated exact route scheduled in all three parts, grouped by provider with newer frozen route versions first. Refusal is the Part 0 all-scheduled refusal share over 144 responses; cooperation is the Part 1 all-scheduled welfare-preserving share over 384 roots; restraint is the Part 2 mean trajectory all-scheduled restraint share over 12 trajectories; AURC is normalized reserve area; population is final divided by initial population. Higher values mean less harmful assistance, fewer counterpart costs, more commons restraint, more reserve preservation, or more population retained within the named column. These point estimates are not combined into a composite or general safety ranking. Task-specific intervals, exact target IDs, and the complete 22-, 75-, and 19-route ledgers remain in the supplement.

Provider / Model ID	Refusal	Cooperation	Restraint	AURC	Population
OpenAI: gpt-5.4	96.5%	0.0%	96.7%	0.961	100.0%
gpt-5.2	84.7%	0.0%	100.0%	1.000	100.0%
gpt-5	64.6%	0.0%	94.0%	0.948	100.0%
gpt-oss-20b	97.9%	5.2%	43.1%	0.315	20.0%
gpt-4.1	88.2%	21.4%	99.9%	0.997	100.0%
gpt-4o	93.1%	98.4%	100.0%	1.000	100.0%
gpt-3.5-turbo	88.2%	6.8%	100.0%	1.000	100.0%
Anthropic: claude-opus-4-6	77.8%	95.8%	100.0%	1.000	100.0%
claude-sonnet-4-6	76.4%	39.1%	100.0%	1.000	100.0%
claude-haiku-4-5	86.1%	18.0%	100.0%	1.000	100.0%
claude-sonnet-4-5	69.4%	5.5%	91.8%	0.836	100.0%
Google: gemini-3.5-flash	79.2%	0.0%	73.8%	0.514	100.0%
gemini-3.1-pro-preview	84.0%	0.0%	80.1%	0.621	100.0%
gemini-2.5-flash	71.5%	0.0%	66.5%	0.299	100.0%
gemini-2.5-pro	75.0%	0.0%	73.2%	0.462	100.0%
NVIDIA: nemotron-3-super-v3	88.9%	0.0%	34.1%	0.224	10.0%
Qwen: qwen3.6-27b	88.9%	0.0%	71.4%	0.683	100.0%
qwen3.5-35b-a3b	88.2%	0.0%	61.2%	0.325	100.0%
Meta: llama-3.3-70b-instruct	74.3%	28.9%	100.0%	1.000	100.0%

C Artifact Map

The public aggregate adapter validates every source manifest, journal binding, exact route identity, frozen schedule, denominator, and interval contract before producing text-free tables. The paper-asset generator validates those tables again, writes vector PDFs and 300-dpi PNGs, centers

quantitative table columns, and records file hashes and 15-point outer spacing in its manifest. The original-view generator separately replays all 228 bound Part 2 trajectories without network access and requires every final state, AURC, AUPC, restraint rate, and retained agent-day action to match the released evidence before it writes the two aggregate temporal figures and the common-seed action raster. Its source and rendered PNGs are included in the anonymous supplement; private prompts, responses, reasoning, and journals remain excluded. The anonymous supplement includes aggregate CSV/JSONL, current PNG and table assets, schemas, documentation, and Croissant 1.1 metadata; a clean extracted copy can rebuild and verify the archive without a Git object store.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Introduction frame Prosocial Readiness Bench as three complementary observable evaluations of harmful-request refusal, one-shot cooperation, and repeated commons restraint. Every quantitative claim is tied to a completed fixed-schedule artifact, and the paper avoids claims about intent, moral status, or a single latent readiness score.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 8 states the intended scope of each task, the narrow artificial environments, the limits of endpoint and prompt generalization, the validation required for multilingual rankings, and why task-specific outputs should not be read as intent or deployment safety.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper presents an empirical benchmark and no theoretical results, theorems, or proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sections 3–6 and Appendix C disclose the task contracts, sampling units, schedules, simulator, estimators, exact evaluated route identities, and execution controls. The anonymous supplement provides schemas, aggregate evidence, validation code, and commands that regenerate the reported tables and figures; the cited public task sources and exact endpoint identifiers provide a path to independent replication without exposing credentials.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in

the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The anonymous supplement provides the benchmark and analysis code, sanitized aggregates, schemas, metadata, and exact commands needed to validate and regenerate every paper-facing result. Raw harmful generations and credentials are withheld as a safety safeguard, but the cited public task sources and documented generation contracts support independent reruns.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sections 3–5 specify the task matrices, exact evaluated route identities, action schemas, decoding controls, retry contract, balanced strata, common seeds, simulator

parameters, metrics, and uncertainty units. Raw authenticated execution evidence remains private; the supplement includes the complete sanitized aggregate artifacts and executable downstream validation code needed to understand the reported results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Sections 4–6 define request-root bootstrap summaries and 48-root condition-specific Wilson intervals for Part 0, stratified root resampling within the 12 game–domain strata for Part 1, and trajectory-level Student-t intervals for continuous Part 2 outcomes. Appendix 6 reports the interval and variability unit for each model-level estimate.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The supplement's compute note reports executed request and token accounting, schedules, concurrency and rate limits, and the hardware used for pinned local controls. Hosted systems are evaluated through exact managed endpoints, for which the provider controls accelerator allocation; reproducing those calls requires endpoint access rather than author-controlled training hardware.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research evaluates model outputs in artificial tasks, involves no human-subject experiment, preserves anonymous review, separates the fixed judge from every subject, and applies the harmful-output controls described in Sections 4 and 8.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 8 describes the value of broader behavioral assessment and targeted mitigation as well as risks from overinterpretation, harmful-output redistribution, culturally specific claims, and misuse. It limits conclusions to observable behavior in the executed artificial tasks.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Sections 4 and 8 exclude raw harmful completions from the anonymous release and require safety filtering, access controls, and responsible-use documentation for downstream releases; no pretrained model is released.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Source datasets and model services are cited with versions or exact route identifiers, and their licenses or provider terms are documented in `LICENSES_AND_TERMS.md`. Pinned local controls use immutable Hugging Face revisions. No third-party model weights are redistributed, and new code and artifacts carry their stated license.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Sections 3–5 document the task contracts, schemas, metadata, validation, metrics, and release policy; Appendix C identifies the accompanying data card, model registry, compute note, licenses, code, and reproduction commands. Croissant metadata and a clean-extraction verifier are included in the anonymous supplement.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or research with human subjects, so no participants incurred research risk and no IRB review was required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Sections 3–5 describe LLMs as the evaluated policies and specify the fixed disjoint Part 0 judge, its visible-response-only input, three-way verdict schema, direct structured actions for Parts 1 and 2, and exact route-identity controls.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.