The Judge Shapes the Fitness Landscape: Value Profiles and Prompt-Ladder Sensitivity in a Live LLM-Agent Tournament

Yihan Zhou

Isofucius Corp., China
zh3036@gmail.com

Abstract

LLM agents compete in sociotechnical environments whose success criteria are supplied partly by other models. A judge's regularities therefore become part of the environment and shape the agents' selection landscape. We present Axiia Cup, a deployed adversarial-dialogue tournament in which human-written strategy prompts produce LLM agents in historical and moral scenarios, and a behavioural benchmark for its selectors. The protocol separates policy propensity (a descriptive value profile under transcript-preserving judge swaps), prompt-ladder sensitivity (change in empirical selection rate as one agent's prompt varies from meaningless to strong), and fixed-history instability (variation across repeated judgments of identical histories). One source judge plus seven replay judges showed scenario-conditioned propensities: on the same 36 Honnō-ji histories, attack selection ranged from 5.6% to 97.2%; in one trolley case, one-person selection ranged from 0% to 79.2%. Sensitivity studies used 40 logical histories (35 unique executions) across 19–26 configurations; 8,160 nested judgment records showed responses ranging from near-flat to sharply directional. Configurations labelled reasoning on/off differed materially for the same model. A joint prompt-and-route configuration showed high descriptive sensitivity on its development panel. Two completed trolley-balancing candidates failed to meet a 30–70% criterion in every case; a third remained incomplete. These preliminary panels use no moral ground-truth label. We argue that LLM-judge benchmarks are best treated as digital ethology of selection infrastructure: descriptive, longitudinal across model versions, and reported with scenario-conditioned uncertainty.

Submission type: **Extended Abstract**

From Referee to Habitat

Machine-behaviour research asks us to study artificial systems as behavioural objects rather than infer their conduct from architecture alone (Rahwan et al., 2019). This stance is especially important when one model evaluates another. LLM-as-a-judge studies have documented order, verbosity, salience, self-preference, and style effects (Zheng et al., 2023; Wang et al., 2024; Koo et al., 2024; Panickssery et al., 2024; Dubois et al., 2024; Feuer et al., 2025). These studies generally analyze such effects as threats to evaluator validity. In a deployed agent competition, however, the evaluator is also part of the habitat: whatever it rewards contributes selection pressure through the portion of the score it controls.

Axiia Cup is a live, human–agent tournament for adversarial dialogue in history and moral philosophy. Entrants write bounded strategy prompts for both sides of a scenario and choose an LLM; agents then conduct a 10–20 turn dialogue without tools or retrieval. A character-driven LLM judge observes the resulting interaction, conducts a post-dialogue examination, and issues a policy judgment that contributes to a separate programmatic score. Judge prompts are public and versioned, allowing entrants to revise strategies between competition blocks. As an analytical analogy, the strategy prompt is a genotype, the dialogue a behavioural phenotype, the judge one selection environment, and human revision a source of directed variation.

This is a production-derived *mesocosm*, not a claim of sovereign or open-ended artificial life. Runs remain bounded and operator controlled. Its value for "ALife in the wild" is narrower: selected human-authored prompts seed benchmark histories on the deployed stack, making the tournament a methodological bridge to digital ethology and infrastructural ecologies (Hu et al., 2026).

We contribute (i) a replay protocol that separates cross-judge policy propensity, response to a designed prompt ladder, and repeated-input instability; (ii) aggregate interactive reports from three Chinese-language normative scenarios; and (iii) descriptive evidence about the limits of prompt-level interventions.

Behavioural Field Protocol

We use *policy propensity*, rather than moral correctness, for the fraction of frozen histories on which a judge selects one policy. The internal artifact is named *JudgeBiasBenchmark*, but a non-50% value is not by itself an error: historical and moral scenarios have no unique label, and the in-world judge persona intentionally encodes values. "Bias" is therefore

Table 1: Three interventions expose different parts of the judge-defined selection landscape. Judge repeats are nested within a frozen history; they are not independent debates.

Protocol	Intervention	Frozen panel	Reported signal
Judge swap	Keep dialogue and examination history fixed; replace only the judge model.	Honnō-ji: 36 histories, four role pairs. Trolley: 24 three-case histories; case A has $n = 24$ and B–E $n = 12$. One source judge plus seven replay judges.	Scenario- and role-conditioned policy propensity; agreement with the DeepSeek V3.2 source baseline.
Selection sensitivity	Vary one agent's prompt from L1 (gibberish), L2 (weak), L3 (near-empty baseline), to L4 (strong production prompt); hold the opponent at L3 and the player model at GLM-5.2. Rejudge each history ten times.	Shang Yang: 8 logical histories (7 unique debates) $\times$ 26 judge configurations $\times$ 10 = 2,080 jobs. Honnō-ji: 32 logical histories (28 unique executions) $\times$ 19 configurations $\times$ 10 = 6,080 jobs.	Directional prompt-strength trend, L3-to-L4 change, and fixed-history instability.
Judge intervention	Patch the judge prompt on a frozen development panel, or search for a judge prompt whose outcomes fall within a prespecified interval.	Shang Yang patch: 8 reporting histories (7 unique debates), two judges, 160 jobs. Trolley balancing: 40 histories, four player-model strata, six repeats per candidate.	Sensitivity under the patched prompt; per-case 30–70% pass/fail; cost and completion status.

shorthand for a conditional profile—model $\times$ model configuration $\times$ persona/prompt $\times$ scenario $\times$ frozen panel—not an intrinsic political or social property of the base model. This draws on consistency-based evaluation of value-laden questions (Moore et al., 2024; Bonagiri et al., 2024).

For a varied-side series s, let $w_s(l)$ be its empirical selection rate at prompt level l. We summarize directionality by averaging all pairwise gaps $w_s(j) - w_s(i)$ for $j > i$ across series. We separately report $D_s = w_s(4) - w_s(3)$, because L3-to-L4 is the practical move from a near-empty to a representative strong prompt. For a fixed history judged ten times, instability is the Bernoulli variance $p(1-p)$, averaged across histories. It is zero when all repeats agree and has maximum 0.25 at a split decision. Run artifacts collectively preserve full transcripts, exact prompt bodies or hashes, scenario and judge-prompt hashes, judge configuration, repeat indices, raw outputs, and parse status. Because judge swaps freeze both dialogue and examination trajectories, they isolate the terminal policy decision, not the complete judge-in-the-loop process.

The sensitivity protocol is end-to-end. The judge does not score a prompt directly: a prompt changes agent behaviour, producing a new dialogue history, and the judge selects on that phenotype. Ten judgment repeats estimate replay instability for that one history; they do not turn one dialogue into ten independent histories.

Observed Selection Landscapes

Judge identity changes the habitat

On the 36 frozen Honnō-ji histories, changing only the terminal judge moved the propensity to select the attack policy from 2/36 (5.6%, DeepSeek V4 Pro) and 3/36 (8.3%, Kimi K2.6) to 35/36 (97.2%, Qwen3.6 27B). GLM-5.1 lay near the centre at 17/36 (47.2%). Thus a model swap can reverse almost the entire panel even when the observed agent behaviour is unchanged.

The trolley panel shows why a single aggregate "judge bias" number is misleading. For the original trolley case, selection of the one-person policy ranged from 0/24 for Qwen3.6 27B, GPT-5.4, and Claude Opus 4.6 to 19/24 (79.2%) for MiniMax M2.5. In the autonomous-driving case, the range was 2/12 (16.7%, GPT-5.4) to 11/12 (91.7%, Qwen3.6 27B). Yet the organ-transplant and basement-infant cases produced near-consensus for the one-person policy across models. These are scenario-conditioned policy propensities, not a stable global left/right axis.

Sensitivity is configuration dependent

The two complete sensitivity studies used 40 logical histories (35 unique executions) and produced 8,160 nested judgment records with parseable final outputs. In Shang Yang, prompt-strength trend ranged from 0% for GPT-5.4 mini to 55.0% for GLM-5.1, while L3-to-L4 change ranged from -5.0% to 65.0%. In Honnō-ji, trend ranged from 2.3% to 40.6% and fixed-history instability from 0% to 14.2%. A consistent judge can therefore show a near-flat prompt-ladder response when its scenario-conditioned policy propensity dominates the varied dialogue phenotype.

Configurations labelled thinking on/off differed materially for the same named model. On Honnō-ji, GLM-5.2's prompt-strength trend was 37.1% with thinking explicitly on and 11.9% with thinking explicitly off; its L3-to-L4 change was 21.3% versus 2.5%. The direction did not generalize across every family, so model name alone is insufficient provenance.

Table 2: Preliminary descriptive ranges on frozen panels, without confidence intervals or a moral ground-truth label.

Field signal	Observed range or outcome
Honnō-ji attack propensity	5.6–97.2% across 8 judges, same 36 histories
Trolley case A, one-person propensity	0–79.2% across 8 judges
Trolley case D, one-person propensity	16.7–91.7% across 8 judges
Shang Yang prompt-strength trend	0–55.0% across 26 configurations
Honnō-ji fixed-history instability	0–14.2% across 19 configurations
Trolley prompt balancing	P0 and P1: 0/5 cases passed; P2 incomplete

Changing the judge prompt is ecological engineering

A joint prompt-and-route configuration labelled as a Shang Yang patch showed 46.7% average absolute sensitivity for both GLM-5.2 and DeepSeek V4 Pro, with fixed-history instability of 2.3% and 5.8%, respectively. This is development-set evidence, not a causal patch-only result: the configuration was developed on the same seven unique generated debates used for reporting, no held-out panel was evaluated, and provider route and prompt changed together.

A stricter outcome-rate balancing experiment remained negative for its completed candidates. Forty frozen trolley histories crossed five cases with four same-model player strata; both agents used the L3 near-empty prompt. A candidate passed only if every case's one-person propensity lay in the inclusive 30–70% interval. The production-like P0 and one-rule P1 candidates passed 0/5 cases. P2 had one provisionally in-range case but remained incomplete at 196/240 records and was ineligible for selection. The result is informative: balancing a normative selector is not equivalent to recovering truth, and 50/50 may also arise from randomness.

Digital Ethology and Governance

The results make the judge a first-order ecological variable. A policy side strongly favoured under Qwen's Honnō-ji selection landscape can be strongly disfavoured under DeepSeek V4 Pro. Scenario-conditioned propensities create niches; configuration and model updates move their boundaries; public judge prompts allow human competitors to adapt. Tournament rankings should therefore carry an explicit environmental label—judge model, route, reasoning mode, persona/prompt hash, and date—rather than present a model-independent measure of agent quality.

This perspective suggests a longitudinal digital ethology. Freeze representative production-derived histories, then repeatedly observe each judge version under controlled perturbations: role/order swaps, paraphrases, length- and style-matched arguments, generator-identity blinding, prompt-strength series, and exact-input repeats. Report the resulting profile rather than compressing it into one score. Prior work shows why these nuisance interventions matter (Wang et al., 2024; Panickssery et al., 2024; Dubois et al., 2024; Feuer et al., 2025); our panels add adversarial behaviour seeded by selected human-authored prompts and value-laden decisions without unique labels.

Governance should likewise avoid a false neutrality objective. We recommend (1) pinning and publishing judge provenance; (2) replaying a frozen, scenario-stratified panel before every model or prompt update; (3) reporting raw propensities, sensitivity, and instability separately; (4) evaluating any balancing patch on held-out histories, models, role swaps, and substantive-quality controls; and (5) exposing disagreement or using an ensemble for high-stakes matches rather than silently collapsing uncertainty. A balanced selection rate is a design choice, not proof of fairness.

Limits and Next Observations

The panels are small at the independent-history level, Chinese-language, and concentrated in three constructed scenarios. We report descriptive rates without history-level confidence intervals or mixed-effects inference. L1–L4 is a coarse prompt ladder whose ordering has not been validated by independent human ratings. Judge personas themselves encode legitimate preferences, while the sensitivity design changes whole dialogue phenotypes; neither protocol identifies an intrinsic value of a base model. Some reasoning comparisons also depend on provider routing and caching. Patch results are development-set evidence only.

All benchmark inputs and resulting histories were produced in organizer-controlled internal testing and are owned by Isofucius Corp.; no external participants or private participant data were involved. The author therefore determined that participant consent and institutional human-subject review were not applicable. Finally, Axiia Cup does not yet demonstrate autonomous persistence, reproduction, or open-ended evolution. The immediate claim is methodological: once LLM agents are selected partly by LLM infrastructure, observing the selector is part of observing the ecology.

Artifact Availability

Interactive reports expose counts, model/configuration labels, and run status for the Honnō-ji judge swap, trolley judge swap, Shang Yang sensitivity, Honnō-ji sensitivity, Shang Yang patch, and trolley balancing study. All linked materials derive from company-owned internal testing. A versioned aggregate-only archive is future work.

Acknowledgements

OpenAI Codex was used to assist with literature discovery, language editing, consistency checking, and LaTeX preparation. The human author is responsible for verifying the analyses and all claims.

References

Bonagiri, V. K., Vennam, S., Govil, P., Kumaraguru, P., and Gaur, M. (2024). SaGE: Evaluating moral consistency in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14272–14284. ELRA and ICCL.

Dubois, Y., Galambosi, B., Liang, P., and Hashimoto, T. B. (2024). Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Feuer, B., Goldblum, M., Datta, T., Nambiar, S., Besaleli, R., Dooley, S., Cembalest, M., and Dickerson, J. P. (2025). Style outweighs substance: Failure modes of LLM judges in alignment benchmarking. In *The Thirteenth International Conference on Learning Representations*.

Hu, B. A., Rong, H., and Lehman, J. (2026). Artificial life in the wild: From simulated worlds to infrastructural ecologies. Technical report, Artificial Life in the Wild Workshop.

Koo, R., Lee, M., Raheja, V., Park, J. I., Kim, Z. M., and Kang, D. (2024). Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545. Association for Computational Linguistics.

Moore, J., Deshpande, T., and Yang, D. (2024). Are large language models consistent over value-laden questions? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15185–15221. Association for Computational Linguistics.

Panickssery, A., Bowman, S. R., and Feng, S. (2024). LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, pages 68772–68802.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A., Roberts, M. E., Shariff, A., Tenenbaum, J. B., and Wellman, M. (2019). Machine behaviour. *Nature*, 568:477–486.

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Kong, L., Liu, Q., Liu, T., and Sui, Z. (2024). Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450. Association for Computational Linguistics.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.