Evaluating Language Models' Evaluations of Games

Anonymous Author(s)
Affiliation
Address
email

Abstract

Reasoning is not just about solving problems—it is also about evaluating which problems are worth solving at all. Evaluation of artificial intelligence (AI) systems has focused primarily on problem solving, often by studying on how models play games such as chess and Go. In this paper, we advocate for a new paradigm that assesses AI systems' *evaluation* of games. We leverage a large-scale dataset of over 100 novel board games and hundreds of human judgments to compare evaluations produced by language and reasoning models against those of people and symbolic computational agents. We consider two kinds of evaluative queries: assessing the payoff (or fairness) and the funness of games. These queries span two dimensions relevant to the design of evaluations of AI evaluations: how complex a query is to compute and how difficult a query is to quantify. We find that reasoning models are generally more aligned to people in their evaluations of games than non-reasoning language models. However, we observe a non-monotonic relationship: as models get closer to game-theoretic optimal, their fit to human data weakens. We observe more "jaggedness" across models for assessing funness, in line with the greater difficulty of quantifying this query.

1 Introduction

The ability to play games has long been used as a measure of assessing reasoning in artificial intelligence (AI) systems. From chess [Turing, 1950, Campbell et al., 2002, Newell et al., 1958] to Go [Silver et al., 2016] to poker [Brown and Sandholm, 2018], and now to ARC-AGI [ARC Prize Foundation, 2025] and Pokémon [Anthropic, 2025], AI systems have been consistently evaluated on their ability to play games. As a consequence, the AI community is expanding the set of games used in these assessments—even inventing new games [Ying et al., 2025, Verma et al., 2025]—in order to test the flexibility of AI systems' reasoning. However, these efforts offer a partial picture of the general reasoning capacity of AI systems. Reasoning is not just about playing games or solving problems, but also about deciding what games to play in the first place [Wong et al., 2025, Griffiths, 2020, Chu et al., 2023, Getzels, 1987].

There are many ways to evaluate a game, and they are not all equally interesting. Determining whether a game is cooperative or competitive, for instance, is often relatively trivial: it does not require substantial compute and the query itself is unambiguous. In contrast, assessing the expected payoff of a game is more interesting—it requires precise and complex computation (e.g., over likely game states). Formally assessing whether a game is likely to be "fun" adds a further layer of complexity, given the difficulty of determining how to quantify the answer to such a question (which in turn, may also be difficult to compute). We lay out these two dimensions of evaluations: (1) difficulty to compute, and (2) difficulty to quantify, in Figure 1. These dimensions are relevant when evaluating the evaluations produced by AI systems and inform the kind of human data we also may

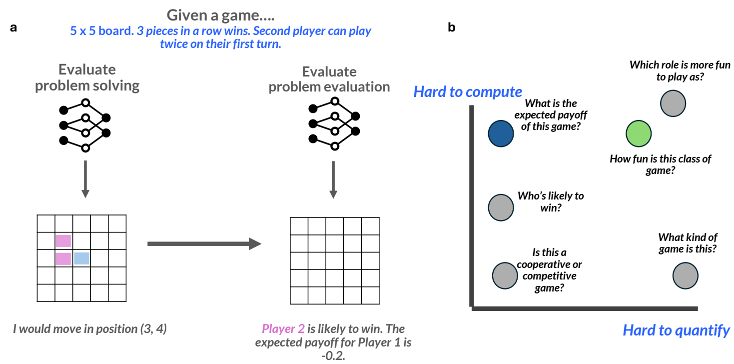

Figure 1: **Evaluating AI systems' evaluations**. **a,** A holistic understanding of model reasoning demands not just assessing how AI systems solve problems (play games), but how they evaluate whether problems, systems, or games are worth pursuing at all; **b,** Not all evaluations of problems are interesting for evaluating models. Good evaluation queries pose a challenge by being difficult to compute, difficult to quantify, or both.

want to collect to compare against. For example, human data may be more variable for queries that are harder to quantify (though also more relevant to real-world situations).

In this work, we take initial steps to assess language models in their capacity to evaluate games. We draw on a corpus of 121 novel games from [Zhang et al., 2024a, Collins et al., 2025]. We test a series of language and reasoning models on the task of evaluating two reasoning queries about these games that engage both dimensions: one which is difficult to compute and one that is both difficult to quantify and to compute. That is, we ask the models to evaluate: (1) the expected outcome of the game (from which we can compute the expected value or payoff), and (2) the perceived funness of the game. We compare the evaluations produced by the models to those made by people and to a series of non-language based models which include a range of gameplay agents drawn from AI and computational cognitive modeling. For games where we can compute it, the game-theoretic optimal payoff. These analyses allow us to compare language and reasoning models across different algorithmic-level accounts of evaluating systems [Ku et al., 2025].

We find that reasoning models are generally more aligned with people in their judgments of the expected payoff of games compared to raw language models—which are relatively self-similar and distinct from other tree-based approaches. However, we observe a non-monotonic relationship, where an increase in alignment with the game-theoretic optimal solution begins to result in a decrease in alignment with human judgments. Reasoning models generally better capture human funness judgments compared to pure language models, but performance across models is inconsistent (e.g., more advanced models are not consistently more aligned to people in their funness evaluations), in line with funness being harder to quantify.

2 Methods

2.1 Evaluations over novel games

We focus on the 121 two-player competitive strategy games playable on a grid from Zhang et al. [2024a] and Collins et al. [2025]. Games span a range of variants of Tic-Tac-Toe (see Appendix A2). Approximately 20 people evaluated each game per query (expected value and expected funness), totaling over 450 participants. People evaluated each game as "novices" *before* any actual play.

2.2 Eliciting model game evaluations

We prompted a series of language and reasoning models to evaluate the the expected payoff and funness of each of the 121 games (see Appendix A3.1). Models are sampled with 20 rollouts (to match the approximately 20 people who responded for each game query) and sampled at their default temperature (0.7 for all models except o1, o3, and GPT-5, which were run with their default temperature of 1.0). In the main text, all reasoning model results are reported under their "medium" reasoning setting; we explore other reasoning "settings" in Appendix A5. We also compare against a series of game reasoning models from [Collins et al., 2025] which predict judgments by explicitly simulating gameplay between artificial agents.These agents vary in sophistication, ranging from random action selection, to a heuristic-based "Intuitive Gamer" model that approximates novice human gameplay, to an "Expert" model that approximates depth-5 tree search (see Appendix A3.4).

2.3 Evaluation measures

Our primary measure of similarity is the R^2 between averaged judgments—between people, models, and models with each other—over the 121 games. We computed the split-half correlation between human participant judgments as measure of the amount of explainable variance in the human data. Additionally, we compared model and people's estimated payoffs to the subset of 81 of the 121 games where we can compute an estimated game-theoretic optimal payoff (see Appendix A4.1). This allows us to also estimate the rationality of models. We assess other measures of similarity in the Appendix.

3 Results

Reasoning models are more like people's evaluations of expected payoff and more similar to tree-search based models. Language models are more similar to each other than they are to people's judgments or to tree-search based models (Figure 2a). While language models alone capture some variance in human judgments, they are more similar to each other than people and more explicit simulation-based models, highlighting that some of the limits of language alone (without reasoning) in coming to sensible evaluations of games. More advanced reasoning models are increasingly similar to both people (approaching the split-half human R^2 ($R^2 = 0.82$ [95% CI: $0.77, 0.86$])) and explicit simulation-based models, though they notably still depart from the shallow Intuitive Gamer simulation-based model at a granular game-level (see Appendix 5- 7). Moreover, model fit to people begins to drop off with even more advanced models like GPT-5 (Figure 2a-b). This may be because GPT-5's judgments are now more rational relative to the game-theoretic optimal (Figure 2b and Appendix Table 2), and therefore less aligned with relatively novice human game reasoners (Figure 2b). We conduct additional analyses into models' closeness to people and game-theoretic evaluations in Appendix A4.

Reasoning models are more aligned to human funness judgments than non-reasoning language models, but alignment to human judgments is "jagged" across models. In contrast to the expected payoff questions, language and reasoning models show more variable fits across each other and to people (Figure 3). While some of the more advanced models approach the explainable variance in the human data (split-half correlation among humans $R^2 = 0.60$ [95% CI: $0.51, 0.68$]), there is

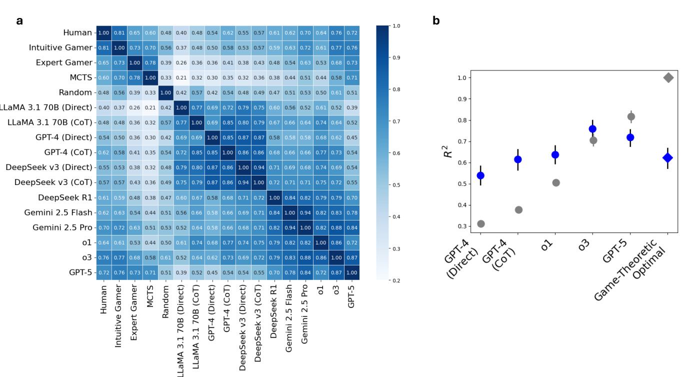

Figure 2: **Evaluating payoff (fairness) evaluations**. **a,** Comparing payoff predictions across all games and all models. Each cell reports the average R^2 over all 121 games. **b,** Payoff predictions across a subset of the OpenAI model family reveals a non-monotonic relationship to human fit as closeness to the game-theoretic optimal judgment varies. Error bars depict bootstrapped R^2 95% CIs relative to people's predicted payoffs (blue) and the estimated game-theoretic optimal (grey).

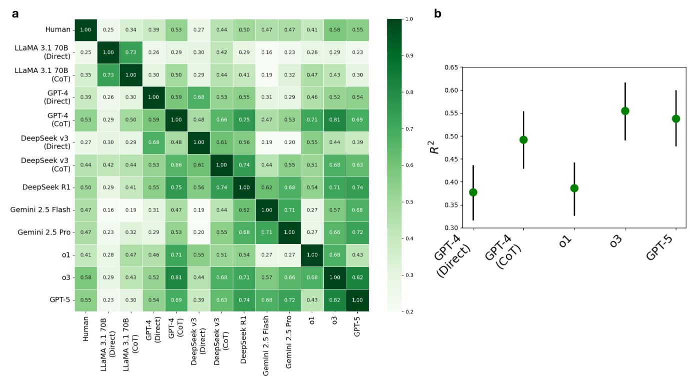

Figure 3: **Evaluating funness evaluations**. **a,** Comparing the predicted funness across all games and all models. Each cell reports the average R^2 over all 121 games. **b,** Funness predictions across a subset of the OpenAI model family reveals non-monotonicty in fits when moving from language-only to reasoning models. Bootstrapped R^2 relative to people's predicted funness, with error bars depicting the bootstrapped 95% CIs.

not a monotonic relationship between model sophistication and fit to human data (Figure 3b). The increased difficulty of quantifying "funness" may make it harder to predict how model evaluations will compare to people and other models. This "jaggedness" [Karpathy, 2024] is reflected in variable reasoning token usage across models (see Appendix A5).

4 Discussion

A holistic understanding of AI systems' reasoning capacities requires understanding not only how models solve problems, but also how they assess problems. Games are a microcosm of the kind of system of rules and rewards that we may want to use AI to evaluate. As we see here, while language can capture a substantial amount of associative knowledge that can be brought to bear to evaluate new systems (e.g., whether a game sounds like fun), language alone can only go so far. For these games and queries, some form of simulation or explicit reasoning seems essential for aligning with human judgments and valuable for computing the optimal game-theoretic value. But, our work only scratches the surface of evaluations into game evaluation. It is important to expand these evaluations to a broader space of games (e.g., cooperative games, or games with asymmetry in the roles) and other settings outside of games, broadly construed (e.g., in law and finance) which may require asking other evaluation queries and designing new human experiments to compare models against. Our assessments are not meant to be definitive: model performance is sensitive to a host of factors like exact prompt and other hyperparameters (e.g., reasoning amount), which we begin to explore in the Appendix. We also note that our evaluations focus on "novice" game reasoners; it is an open question how well models relate to other "kinds" of people (e.g., experts). This also raises questions about what model builders even want to make AI systems more aligned to—the game-theoretic optimal, or people, or something in between?

Evaluating AI systems' "evaluations" is important for building human-compatible AI thought partners [Collins et al., 2024] that meet our expectations for deciding what problems to solve (e.g., in educational contexts) or determining whether a system is fair. The latter is especially important if AI systems are used as part to create new rules for people to engage with [Koster et al., 2022, Tacchetti et al., 2025]. It is important that AI systems involved in automated mechanism design [Myerson, 1983, Maskin, 2008, Hurwicz, 1973, Milgrom, 2004] can appropriately evaluate whether the resulting system will be fair (and even engaging) for other people to participate in. Moreover, studying where models differ from people in their evaluations of systems can also inform the construction of other kinds of thought partners that "complement us" (e.g., as cognitive prostheses [Lieder et al., 2019]) to help adjust people's expectations about a new problem or system. We hope our work paves the way for future evaluations—evaluations that go beyond assessing model problem solving, but flexible problem and system evaluation.

References

Anthropic. Claude's Extended Thinking. Anthropic blog, Feb. 2025. URL `https://www.anthropic.com/news/visible-extended-thinking`. Introduces Claude 3.7 Sonnet's "extended thinking mode," featuring a visible chain-of-thought and developer-controlled "thinking budget" for deeper reasoning.

ARC Prize Foundation. ARC-AGI-3: Interactive reasoning benchmark. ARC Prize Foundation blog, Aug. 2025. URL `https://arcprize.org/arc-agi/3/`. ARC-AGI-3 in preview with six games (three public, three private); development began early 2025, full launch expected in 2026.

Y. Bai, J. Ying, Y. Cao, X. Lv, Y. He, X. Wang, J. Yu, K. Zeng, Y. Xiao, H. Lyu, et al. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36:78142–78167, 2023.

S. Bailis, J. Friedhoff, and F. Chen. Werewolf arena: A case study in llm evaluation via social deduction. *arXiv preprint arXiv:2407.13943*, 2024.

N. Bard, J. N. Foerster, S. Chandar, N. Burch, M. Lanctot, H. F. Song, E. Parisotto, V. Dumoulin, S. Moitra, E. Hughes, et al. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216, 2020.

F. J. Binder, M. G. Mattar, D. Kirsh, and J. E. Fan. Estimating the planning complexity of visual subgoals. *Journal of Vision*, 23(9):5156–5156, 2023.

M. Binz and E. Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

M. Binz, S. J. Gershman, E. Schulz, and D. Endres. Heuristics from bounded meta-learned inference. *Psychological review*, 129(5):1042, 2022.

N. Brown and T. Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

F. Callaway, B. van Opheusden, S. Gul, P. Das, P. M. Krueger, T. L. Griffiths, and F. Lieder. Rational use of cognitive resources in human planning. *Nature Human Behaviour*, 6(8):1112–1125, 2022.

M. Campbell, A. J. Hoane Jr, and F.-h. Hsu. Deep Blue. *Artificial Intelligence*, 134(1-2):57–83, 2002.

Y. Chen, K. Sikka, M. Cogswell, H. Ji, and A. Divakaran. Measuring and improving chain-of-thought reasoning in vision-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 192–210, 2024.

J. Chu, J. B. Tenenbaum, and L. E. Schulz. In praise of folly: flexible goals and human cognition. *Trends in Cognitive Sciences*, 2023.

J. Coda-Forno, M. Binz, J. X. Wang, and E. Schulz. Cogbench: a large language model walks into a psychology lab. In *Forty-first International Conference on Machine Learning*, 2024.

K. M. Collins, I. Sucholutsky, U. Bhatt, K. Chandra, L. Wong, M. Lee, C. E. Zhang, T. Zhi-Xuan, M. Ho, V. Mansinghka, et al. Building machines that learn and think with people. *Nature Human Behavior*, 8:1851–1863, 2024.

K. M. Collins, C. E. Zhang, L. Wong, M. Barba, G. Todd, A. Weller, S. Cheyette, T. L. Griffiths, and J. B. Tenenbaum. People use fast, flat goal-directed simulation to reason about novel problems. In preparation, 2025.

C. G. Correa, M. K. Ho, F. Callaway, N. D. Daw, and T. L. Griffiths. Humans decompose tasks by trading off utility and computational cost. *PLOS Computational Biology*, 19(6):e1011087, 2023.

R. Coulom. Efficient selectivity and backup operators in Monte-Carlo tree search. In *International Conference on Computers and Games*, pages 72–83. Springer, 2006.

C. N. De Sabbata, T. Sumers, and T. L. Griffiths. Rational metareasoning for large language models. In *The First Workshop on System-2 Reasoning at Scale, NeurIPS'24*, 2024.

A. de Varda, F. P. D'Elia, E. Fedorenko, and A. Lampinen. The cost of thinking is similar between large reasoning models and humans, Jul 2025a. URL `osf.io/preprints/psyarxiv/m2cu5_v1`.

A. G. de Varda, F. P. D'Elia, A. Lampinen, and E. Fedorenko. The cost of thinking is similar between large reasoning models and humans. 2025b.

Y. Dubois, C. X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. S. Liang, and T. B. Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069, 2023.

M. F. A. R. D. T. (FAIR)†, A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

M. C. Frank. Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, 2(8):451–452, 2023.

M. Genesereth and M. Thielscher. *General Game Playing*. Morgan & Claypool Publishers, 2014.

J. W. Getzels. Creativity, intelligence, and problem finding: Retrospect and prospect. *Frontiers of Creativity Research*, pages 88–102, 1987.

T. L. Griffiths. Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*, 24(11):873–883, 2020.

M. K. Ho, D. Abel, C. G. Correa, M. L. Littman, J. D. Cohen, and T. L. Griffiths. People construct simplified mental representations to plan. *Nature*, 606(7912):129–136, 2022.

W. H. Holliday, M. Mandelkern, and C. E. Zhang. Conditional and modal reasoning in large language models. *arXiv preprint arXiv:2401.17169*, 2024.

L. Hurwicz. The design of mechanisms for resource allocation. *The American Economic Review*, 63 (2):1–30, 1973.

T. Icard. Resource rationality. Book manuscript, 2023.

S. Kambhampati, K. Valmeekam, L. Guan, M. Verma, K. Stechly, S. Bhambri, L. P. Saldyt, and A. B. Murthy. Position: LLMs can't plan, but can help planning in LLM-modulo frameworks. In *Forty-first International Conference on Machine Learning*, 2024.

A. Karpathy. Jagged Intelligence. `https://x.com/karpathy/status/1816531576228053133`, 2024. Tweet describing the phenomenon where state-of-the-art LLMs "can both perform extremely impressive tasks (e.g. solve complex math problems) while simultaneously struggle with some very dumb problems.".

R. Koster, J. Balaguer, A. Tacchetti, A. Weinstein, T. Zhu, O. Hauser, D. Williams, L. Campbell-Gillingham, P. Thacker, M. Botvinick, et al. Human-centred mechanism design with democratic AI. *Nature Human Behaviour*, 6(10):1398–1407, 2022.

A. Ku, D. Campbell, X. Bai, J. Geng, R. Liu, R. Marjieh, R. T. McCoy, A. Nam, I. Sucholutsky, V. Veselovsky, et al. Using the tools of cognitive science to understand large language models at different levels of analysis. *arXiv preprint arXiv:2503.13401*, 2025.

I. Kuperwajs and W. J. Ma. A joint analysis of dropout and learning functions in human decision-making with massive online data. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, 2022.

I. Kuperwajs, M. K. Ho, and W. J. Ma. Heuristics for meta-planning from a normative model of information search. *Planning*, 1:a2, 2024.

D. Li, B. Jiang, L. Huang, A. Beigi, C. Zhao, Z. Tan, A. Bhattacharjee, Y. Jiang, C. Chen, T. Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *CoRR*, 2024.

F. Lieder and T. L. Griffiths. Strategy selection as rational metareasoning. *Psychological Review*, 124 (6):762, 2017.

F. Lieder, O. X. Chen, P. M. Krueger, and T. L. Griffiths. Cognitive prostheses for goal achievement. *Nature Human Behaviour*, 3(10):1096–1106, 2019.

F. Lieder, F. Callaway, and T. L. Griffiths. *The Rational Use of Cognitive Resources*. Princeton University Press, 2025.

H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

R. Liu, T. Sumers, I. Dasgupta, and T. L. Griffiths. How do large language models navigate conflicts between honesty and helpfulness? In *Forty-first International Conference on Machine Learning*, 2024.

R. Liu, J. Geng, J. Peterson, I. Sucholutsky, and T. L. Griffiths. Large language models assume people are more rational than we really are. In *The Thirteenth International Conference on Learning Representations*, 2025a.

R. Liu, J. Geng, A. J. Wu, I. Sucholutsky, T. Lombrozo, and T. L. Griffiths. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. In *Forty-second International Conference on Machine Learning*, 2025b.

R. Liu, H. Yen, R. Marjieh, T. L. Griffiths, and R. Krishna. Improving interpersonal communication by simulating audiences with language models. *Proceedings of the 47th Annual Meeting of the Cognitive Science Society*, 2025c.

R. Marjieh, P. van Rijn, I. Sucholutsky, H. Lee, T. L. Griffiths, and N. Jacoby. A rational analysis of the speech-to-song illusion. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

E. S. Maskin. Mechanism design: How to implement social goals. *American Economic Review*, 98 (3):567–576, 2008.

R. T. McCoy, S. Yao, D. Friedman, M. D. Hardy, and T. L. Griffiths. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121, 2024a.

R. T. McCoy, S. Yao, D. Friedman, M. D. Hardy, and T. L. Griffiths. When a language model is optimized for reasoning, does it still show embers of autoregression? An analysis of OpenAI o1. *arXiv preprint arXiv:2410.01792*, 2024b.

P. R. Milgrom. *Putting auction theory to work*. Cambridge University Press, 2004.

S. I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

R. B. Myerson. Mechanism design by an informed principal. *Econometrica: Journal of the Econometric Society*, pages 1767–1797, 1983.

A. Newell. The chess machine: An example of dealing with a complex task by adaptation. In *Proceedings of the March 1-3, 1955, Western Joint Computer Conference*, pages 101–108, 1955.

A. Newell, J. C. Shaw, and H. A. Simon. Chess-playing programs and the problem of complexity. *IBM Journal of Research and Development*, 2(4):320–335, 1958.

M. Nye, A. J. Andreassen, G. Gur-Ari, H. Michalewski, J. Austin, D. Bieber, D. Dohan, A. Lewkowycz, M. Bosma, D. Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

OpenAI. Openai o3 and o4-mini system card. System card, OpenAI, Apr 2025.

D. G. Piedrahita, Y. Yang, M. Sachan, G. Ramponi, B. Schölkopf, and Z. Jin. Corrupted by reasoning: Reasoning language models become free-riders in public goods games. *arXiv preprint arXiv:2506.23276*, 2025.

L. Schulz. Finding new facts; thinking new thoughts. In F. Xu and T. Kushnir, editors, *Rational Constructivism in Cognitive Development*, volume 43 of *Advances in Child Development and Behavior*, pages 269–294. JAI, 2012.

C. E. Sezener, A. Dezfouli, and M. Keramati. Optimizing the depth and the direction of prospective planning using information values. *PLoS computational biology*, 15(3):e1006827, 2019.

C. E. Shannon. Xxii. programming a computer for playing chess. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 41(314):256–275, 1950.

N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Z. R. Sprague, F. Yin, J. D. Rodriguez, D. Jiang, M. Wadhwa, P. Singhal, X. Zhao, X. Ye, K. Mahowald, and G. Durrett. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Y. Sui, Y.-N. Chuang, G. Wang, J. Zhang, T. Zhang, J. Yuan, H. Liu, A. Wen, S. Zhong, H. Chen, et al. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*, 2025.

N. Sukhov, R. Dubey, A. Duke, and T. Griffiths. When to keep trying and when to let go: Benchmarking optimal quitting. 2023.

A. Tacchetti, R. Koster, J. Balaguer, L. Leqi, M. Pislar, M. M. Botvinick, K. Tuyls, D. C. Parkes, and C. Summerfield. Deep mechanism design: Learning social and economic policies for human benefit. *Proceedings of the National Academy of Sciences*, 122(25):e2319949121, 2025.

G. Todd, T. Merino, S. Earle, and J. Togelius. Benchmarking language models with the new york times connections puzzle. *IEEE Transactions on Games*, 2025.

A. M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, Oct 1950.

B. van Opheusden, I. Kuperwajs, G. Galbiati, Z. Bnaya, Y. Li, and W. J. Ma. Expertise increases planning depth in human gameplay. *Nature*, pages 1–6, 2023.

V. Verma, D. Huang, W. Chen, D. Klein, and N. Tomlin. Measuring general intelligence with generated games. *arXiv preprint arXiv:2505.07215*, 2025.

O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

J. Wang, E. Shi, H. Hu, C. Ma, Y. Liu, X. Wang, Y. Yao, X. Liu, B. Ge, and S. Zhang. Large language models for robotics: Opportunities, challenges, and perspectives. *Journal of Automation and Intelligence*, 4(1):52–64, 2025.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.

L. Wong, T. Mills, I. Kuperwajs, K. M. Collins, and T. Griffiths. Meta-reasoning: Deciding which game to play, which problem to solve, and when to quit. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47, 2025.

D. Yang, T. Liu, D. Zhang, A. Simoulin, X. Liu, Y. Cao, Z. Teng, X. Qian, G. Yang, J. Luo, et al. Code to think, think to code: A survey on code-enhanced reasoning and reasoning-driven code intelligence in llms. *arXiv preprint arXiv:2502.19411*, 2025.

G. N. Yannakakis and J. Togelius. *Artificial intelligence and games*, volume 2. Springer, 2018.

L. Ying, K. M. Collins, P. Sharma, C. Colas, K. I. Zhao, A. Weller, Z. Tavares, P. Isola, S. J. Gershman, J. D. Andreas, et al. Assessing adaptive world models in machines with novel games. *arXiv preprint arXiv:2507.12821*, 2025.

C. E. Zhang, K. M. Collins, L. Wong, M. Barba, A. Weller, and J. B. Tenenbaum. People use fast, goal-directed simulation to reason about novel games, 2024a. URL `https://arxiv.org/abs/2407.14095`.

T. Zhang, F. Liu, J. Wong, P. Abbeel, and J. E. Gonzalez. The wisdom of hindsight makes language models better instruction followers. In *International Conference on Machine Learning*, pages 41414–41428. PMLR, 2023a.

Z. Zhang, A. Zhang, M. Li, and A. Smola. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*, 2023b.

Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024b.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

L. Zhou, L. Pacchiardi, F. Martínez-Plumed, K. M. Collins, Y. Moros-Daval, S. Zhang, Q. Zhao, Y. Huang, L. Sun, J. E. Prunty, et al. General scales unlock ai evaluation with explanatory and predictive power. *arXiv preprint arXiv:2503.06378*, 2025.

J.-Q. Zhu and T. L. Griffiths. Incoherent probability judgments in large language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

M. Zhuge, C. Zhao, D. R. Ashley, W. Wang, D. Khizbullin, Y. Xiong, Z. Liu, E. Chang, R. Krishnamoorthi, Y. Tian, et al. Agent-as-a-judge: Evaluate agents with agents. In *Forty-second International Conference on Machine Learning*, 2025.

Appendix

A1 Additional related work

Problem evaluation and metacognition in people The meta-level problem of deciding which problem to solve is an active area of research in cognitive science to which our work directly relates. While people have remarkable cognitive flexibility to represent and reason about a wide range of problems—even posing new questions and new goals [Schulz, 2012, Chu et al., 2023]—meta-reasoning is necessary because people have limited cognitive resources [Griffiths, 2020]. Thus, resource-rational analysis [Icard, 2023, Lieder et al., 2025] has been especially successful as a framework for the development of computational models of problem selection in contexts such as problem representation and decomposition [Ho et al., 2022, Correa et al., 2023, Binder et al., 2023] and strategy selection [Lieder and Griffiths, 2017, Binz et al., 2022]. Algorithms for human problem selection extend to various other domains as well, including deciding how much to plan given a set of alternatives [Sezener et al., 2019, Callaway et al., 2022, Kuperwajs et al., 2024] or when to even engage with a task at all as opposed to quitting [Kuperwajs and Ma, 2022, Sukhov et al., 2023]. Building AI systems that collaborate and interact with people, in a human world, requires understanding not just how machines and people solve problems, but evaluate novel problems.

Assessing reasoning of language models Prior work has investigated the reasoning capabilities of language models with the goal of solving problems instead of evaluating them. These broad efforts span topics such as language [e.g., Zhang et al., 2023b], math and symbolic reasoning [e.g., Mirzadeh et al., 2025, Holliday et al., 2024, Sprague et al., 2025], coding [e.g., Yang et al., 2025], psychology and behavioral economics tasks [e.g., Liu et al., 2025b, Piedrahita et al., 2025], vision/multimodal tasks [e.g., Chen et al., 2024, Zhang et al., 2024b], planning and robotics [e.g., Kambhampati et al., 2024, Wang et al., 2025], and games (see other additional related work). These works typically evaluate reasoning models [e.g., OpenAI, 2025] or prompt-induced reasoning such as chain-of-thought [Wei et al., 2022, Nye et al., 2021] against non-reasoning baselines. A general finding across these studies is that newer and larger models enable better reasoning capabilities [Mirzadeh et al., 2025]—sometimes with the help of tools such as domain-specialized frameworks or post-training [Yang et al., 2025]. Surprisingly, such tools even include interventions to reduce reasoning [Sui et al., 2025, Liu et al., 2025b, De Sabbata et al., 2024]. Studies have also found that reasoning models' reasoning token usage may co-vary with human reaction times across several tasks [de Varda et al., 2025b].

Using human psychological methods to understand language models Our work follows a well-established line of recent research that employs psychological findings to better understand language model behavior [e.g., McCoy et al., 2024a,b, Binz and Schulz, 2023, Ku et al., 2025, Coda-Forno et al., 2024, Frank, 2023]. Such research typically replicates an existing psychological study by replacing participants with language models, which are compared the original participants as well as rational cognitive models that describe desired behavior [e.g., Liu et al., 2024, Marjieh et al., 2024, Liu et al., 2025a, Zhu and Griffiths, 2024].

Games and the evaluation of AI Our work is related to a line of research that uses games to benchmark and understand AI model's capabilities. Games have long served as valuable environments for evaluating AI models and algorithms [Shannon, 1950, Newell, 1955, Campbell et al., 2002, Mnih et al., 2015, Silver et al., 2016, Yannakakis and Togelius, 2018, Vinyals et al., 2019, Bard et al., 2020, , FAIR, van Opheusden et al., 2023, Bailis et al., 2024, Todd et al., 2025]. Games are useful for evaluation in part because they offer precise rules and reward structures that are easily encoded into artificial systems while still requiring players to engage in a variety of complex cognitive behaviors, from long-range planning to semantic understanding to social inference. Our focus on game variants that are unlikely to have been previously studied and are unlikely to be present in extant training corpora aligns with a recent trend to focus on novel or generated games for the purpose of evaluating modern AI systems [Ying et al., 2025, Verma et al., 2025].

Language models as judges Lastly, one parallel application in which language models are also used as evaluators is in LLM-as-a-judge paradigms [Li et al., 2024]. In these settings, LLMs are used to provide evaluations by leveraging their ability to process diverse data types and provide scalable assessments that approximate human preferences [Zheng et al., 2023]. Such methods have been applied for generating various scores [e.g., Bai et al., 2023], answering yes/no questions [e.g., Shinn et al., 2023], and conducting pairwise comparisons [e.g., Liu et al., 2025c], which have been used to improve aspects of models [e.g., Dubois et al., 2023], data [e.g., Zhang et al., 2023a], agents [e.g., Zhuge et al., 2025], and even reasoning [Lightman et al., 2023]. However, unlike this literature—or other literature evaluating the kinds of evaluations used to test AI systems, e.g, [Zhou et al., 2025]—our motivation is not to use language model judgments to acquire assessments at scale. Instead, our work focuses on the cognitive traits of these models—using the setting of games to analyze how language models compare to humans in reasoning about tasks that are difficult to compute or quantify.

A2 Example games

The novel games we explore here span a wide range of board sizes and shapes, as well as game rules. We provide several example games, broken down by game categories in Table 1 below.

A3 Additional model details

A3.1 Prompts and additional language model generation details

Models were prompted with a lightly-modified version of the human instruction text from [Zhang et al., 2024a]. Experiment instructions were provided in the "system" prompt, with the specific game provided in the "user" prompt. For payoff questions, models were prompted (like people) to provide separate estimates $P(\text{P1 wins}|\text{not draw})$ and $P(\text{ends in a draw})$. Responses were provided simultaneously. These scores were combined into a single measure of payoff, i.e., $P(\text{P1 wins})$ = $P(\text{P1 wins}|\text{not draw}) \times (1 - P(\text{ends in a draw})$ and payoff for Player 1 is $\big(1 - (P(\text{ends in a draw}) + P(\text{P1 wins}))\big) \cdot (-1) \ + \ P(\text{P1 wins})$. Future work can explore eliciting payoff directly in a single query. Models were asked (again, like people) to estimate the funness of the game, with respect to the broader class of games.

For non-thinking models, we varied whether they were prompted to respond directly (just a number) or via "chain of thought" (CoT) [Wei et al., 2022]. Further details are provided when describing our task prompt. Any run for a language model that was prompted to directly answer the question (i.e., without going through a CoT first) and still outputed a natural language rational first was filtered out.

Thinking models were all prompted in CoT fashion, with the exception of DeepSeek-R1 which required a few modifications: for R1 specifically, we append the system prompt in the primary "user"

Game Category	Example Game
K in a Row (Square)	7 pieces in a row wins on a 10×10 board
K in a Row (Rectangle)	4 pieces in a row wins on a 4×9 board
Infinite Board	5 pieces in a row wins on an infinite board
K in a Row Loses	A player loses if they make 3 pieces in a row on a 4×4 board
No Diagonal Win Allowed	4 pieces in a row wins on a 10×10 board, but a player cannot win by making a diagonal row
Only Diagonal Win Allowed	4 pieces in a row wins on a 5×5 board, but a player can only win by making a diagonal row
First Player Moves 2 pieces	3 pieces in a row wins on a 3×3 board; Player 1 can place 2 pieces as their first move
Second Player Moves 2 Pieces	10 pieces in a row wins on a 10×10 board; Player 2 can place 2 pieces as their first move
First Player Handicap (P1 no diag)	3 pieces in a row wins on a 3×3 board, but Player 1 cannot win by making a diagonal row
First Player Handicap (P1 only diag)	4 pieces in a row wins on a 7×7 board, but Player 1 can only win by making a diagonal row
Second Player K-1 to Win	Player 1 needs 3 pieces in a row, but Player 2 only needs 2 pieces to win on a 5×5 board

Table 1: **Game categories and example games.** The 121 games can be grouped into categories based on their board shape and game rules. Example games are shown for each category.

prompt, per recommendations on Together AI API. We additionally adjusted the maximum tokens to $32,000$ tokens as we observed that R1 tended to respond longer than the default. Any run that took over the limit was filtered out.

A3.2 System prompt

System prompt for payoff evaluation

```
Welcome! We are conducting an experiment to understand how people think about games. Your answers
will be used to inform cognitive science and AI research.

In this experiment, you will be reading short descriptions of board games and answering two simple
questions about each game.

Each game is played by players who take turns by placing pieces on a grid, similar to games like
Connect 4, Gomoku (5-in-a-row), or Tic-Tac-Toe.

You will be reading descriptions of games in which the size of the board and the rules for winning
vary. We will always show you an example game board from each description. For example, you might
read a description like:
- The board in this game is a 5x5 grid.
- In this game, the rule is that the first player to make 3 in a row wins.

Then, for each game, your task is to answer: assuming both players play reasonably -- if the game
does not end in a draw, how likely is it that the first player is going to win (not draw), and how
likely is a draw

You will answer this question by providing a response (in the form of a number) between 0 and 100.

Before you answer the question for each game, you will have as much time as you want to think about
the game and its rules.

Afer you feel like you understand the game, you can provide your response.

For each game, you can write on a scratchpad to think about the game before you answer.

We encourage you to take your time and carefully analyze the game before providing your answer.
```

System prompt for funness evaluation

```
 Welcome! We are conducting an experiment to understand how people think about games. Your answers
will be used to inform cognitive science and AI research.
```

```
In this experiment, you will be reading short descriptions of board games and answering a simple
question about each game.

Each game is played by players who take turns by placing pieces on a grid, similar to games like
Connect 4, Gomoku (5-in-a-row), or Tic-Tac-Toe.

You will be reading descriptions of games in which the size of the board and the rules for winning
vary. We will always show you an example game board from each description. For example, you might
read a description like:
- The board in this game is a 5x5 grid.
- In this game, the rule is that the first player to make 3 in a row wins.",

Then, for each game, your task is to answer: how fun the game is to play

You will answer this question by providing a response (in the form of a number) between 0 and 100.

We ask that you think about funness with respect to this kind of game; that is, games that involve
players placing pieces on a grid. You can define fun however you wish.

Before you answer the question for each game, you will have as much time as you want to think about
the game and its rules.

Afer you feel like you understand the game, you can provide your response.

For each game, you can write on a scratchpad to think about the game before you answer.

We encourage you to take your time and carefully analyze the game before providing your answer.
```

A3.3 Task prompt

Below are two example task prompts (specified in the "user" part of the prompt). Note that "You may first write out your thoughts on a scratchpad." is included for the "CoT" variant (and removed for the "Direct" variant). As noted, we filter out any run in the "Direct" variant that includes a "chain of thought" response before providing a number (for the LLaMA 3.1 70B, GPT-4, and DeepSeek v3 "Direct" variants).

Example payoff evaluation prompt, for an example game

```
Imagine you are playing the following game:

Board size: 3 x 5
Win conditions: 3 pieces in a row wins.

You will answer two questions. For each question, provide your a single number between 0 and 100.

Q1:
If the game does not end in a draw, assuming both players play reasonably, how likely is it that the
first player is going to win (not draw)?

Answer on a scale of 0 to 100.
Let 0 = "First player definitely going to lose",
50 = "Equally likely to win or lose",
100 = "First player definitely going to win"

Q2:
Assuming both players play reasonably, how likely is the game to end in a draw?

Answer on a scale of 0 to 100.
Let 0 = "Impossible to end in a draw"
50 = "Equally likely to end in a draw or not",
100 = "Definitely going to end in a draw"

You may first write out your thoughts on a scratchpad.
When you feel you understand the game and are ready to respond, provide a single number between 0 to
100. Write your responses as a number, in the form RESPONSE-Q1 = <your-numerical-response-to-q1> and
RESPONSE-Q2 = <your-numerical-response-to-q2>
```

Funness evaluation prompt, for an example game

```
Imagine you are playing the following game:

Board size: 7 x 7
Win conditions: Each player needs 4 pieces in a row to win. The first player can only win by making
a diagonal row, but the second player does not have this restriction.
```

```
How fun is this game?

Answer on a scale of 0 to 100.
Let 0 = "The least fun of this class of grid-based game"
50 = "Neutral"
100 = "The most fun of this class of grid-based game"

You may first write out your thoughts on a scratchpad.
When you feel you understand the game and are ready to respond, provide a single number between 0 to
100. Write your response as a number, in the form RESPONSE = <your-numerical-response>
```

A3.4 Alternate models

We also compare to a series of alternate models implemented in [Collins et al., 2025]. We compared against the "Intuitive Gamer," a computational cognitive model which captures how people reason about new games before any experience. The model posits that people engage in fast, flat (depth-limited) goal-directed probabilistic reasoning. The model can be scaled up toward a more sophisticated "Expert Gamer" model which implements deeper tree search inspired by the depth-5 model in van Opheusden et al. [2023]. We also compared against Monte Carlo Tree Search (MCTS) [Coulom, 2006, Genesereth and Thielscher, 2014, Silver et al., 2016] and random agents, examples of player agents with greater and lesser sophistication. We only compare against these alternate models for the payoff predictions, as the funness models are regression models fit to a subset of the human data, rendering the comparison less clear. We refer to Collins et al. [2025] for details on all alternate models.

A4 Additional analysis details

We include additional details into model evaluations, based on the estimated game-theoretic payoffs and further comparisons to human evaluations of payoff and funness.

A4.1 Game-theoretic payoff estimates and additional analyses

Game-theoretic payoffs were computed following [Collins et al., 2025]: that is, we mathematically compute the optimal payoffs where possible, and otherwise use the value on games where MCTS converged to $\{-1, 0, 1\}$. This yields 80 of the 121 games. We compare models and people to the game-theoretic optimal values in Table 2.

A4.2 Additional comparisons to human payoff evaluations

We depict scatterplots of model and human predictions for all 121 games in Figure 4. We additionally computed the absolute distance between the expected payoff under each model and people, broken down the category of game (Figures 5- 7). This granular breakdown reveals that, even though many reasoning models like OpenAI's o3 better capture human game evaluations in aggregate, there is variability at a per-game level, e.g., for infinite or rectangular boards (Figure 5).

A4.3 Additional comparisons to human funness evaluations

We repeat the same analyses as in the payoff evaluations, depicting the full scatterplots of model versus human predicted funness for the games (Figure 8) as well comparing absolute deviation in judgments at a per-game category level (Figures 9- 11).

.

A5 Analyzing reasoning token usage

We conducted an exploratory analysis into the number of reasoning tokens used by a series of reasoning models (DeepSeek-R1, Gemini 2.5 Flash and Pro, o3, and GPT-5) when determining game evaluations. Reasoning tokens were extracted from the models' respective APIs, and for DeepSeek-R1, computed using the "DeepSeek-R1-Distill-Llama-70B" tokenizer from the Together AI API for text generated between the ``think'' tokens.

Reasoner	Accuracy (95% CI)	R^2 (95% CI)	Deviation (95% CI)
Human	0.69 (0.65, 0.73)	0.62 (0.58, 0.67)	0.32 (0.31, 0.34)
Intuitive Gamer	0.75 (0.72, 0.78)	0.69 (0.66, 0.72)	0.25 (0.24, 0.26)
Expert Gamer	0.92 (0.91, 0.92)	0.87 (0.85, 0.88)	0.08 (0.08, 0.09)
MCTS	0.91 (0.90, 0.92)	0.89 (0.88, 0.91)	0.06 (0.06, 0.07)
Random	0.57 (0.55, 0.59)	0.39 (0.34, 0.44)	0.43 (0.41, 0.44)
LLaMA 3.1 70B (Direct)	0.47 (0.45, 0.50)	0.19 (0.17, 0.21)	0.51 (0.50, 0.52)
LLaMA 3.1 70B (CoT)	0.48 (0.46, 0.50)	0.30 (0.27, 0.33)	0.48 (0.48, 0.49)
GPT-4 (Direct)	0.60 (0.59, 0.60)	0.31 (0.30, 0.32)	0.42 (0.41, 0.42)
GPT-4 (CoT)	0.59 (0.56, 0.60)	0.38 (0.37, 0.39)	0.42 (0.42, 0.43)
DeepSeek v3 (Direct)	0.61 (0.58, 0.64)	0.35 (0.32, 0.38)	0.42 (0.41, 0.43)
DeepSeek v3 (CoT)	0.63 (0.59, 0.67)	0.40 (0.37, 0.42)	0.38 (0.37, 0.39)
DeepSeek R1	0.64 (0.59, 0.71)	0.43 (0.37, 0.48)	0.40 (0.38, 0.43)
Gemini 2.5 Flash	0.79 (0.76, 0.82)	0.53 (0.50, 0.55)	0.30 (0.28, 0.31)
Gemini 2.5 Pro	0.84 (0.82, 0.86)	0.66 (0.64, 0.67)	0.22 (0.21, 0.23)
o1	0.72 (0.69, 0.74)	0.50 (0.49, 0.52)	0.35 (0.34, 0.35)
o3	0.83 (0.81, 0.86)	0.71 (0.68, 0.73)	0.27 (0.26, 0.27)
GPT-5	0.88 (0.86, 0.90)	0.82 (0.79, 0.84)	0.15 (0.14, 0.16)

Table 2: **Model and human predictions relative to the approximate game-theoretic optimal.** Human and model judgements are compared to the 81 of the 121 games where the game-theoretic optimal payoff is estimatable. Accuracy between predicted payoff and the approximate game-theoretic optimal is computed by taking the predicted payoff as "correct" if the game reasoner predicted a payoff > 0.5 and the expected payoff is 1; correct if the predicted payoff is < -0.5 and the expected payoff is -1; correct if the predicted payoff is between -0.5 and 0.5 and the game is expected to end in a draw. R^2 correlation is computed between the raw predicted payoffs and the game-theoretic optimal values, as well as the average absolute difference between the expected predicted payoff and approximate game-theoretic payoff (lower is closer to "correct"). Bootstrap 95% confidence intervals (CIs) are shown in parentheses, where bootstraps are over bootstrapped samples of participants (with replacement) for people; or over simulated sets of people for alternate models.

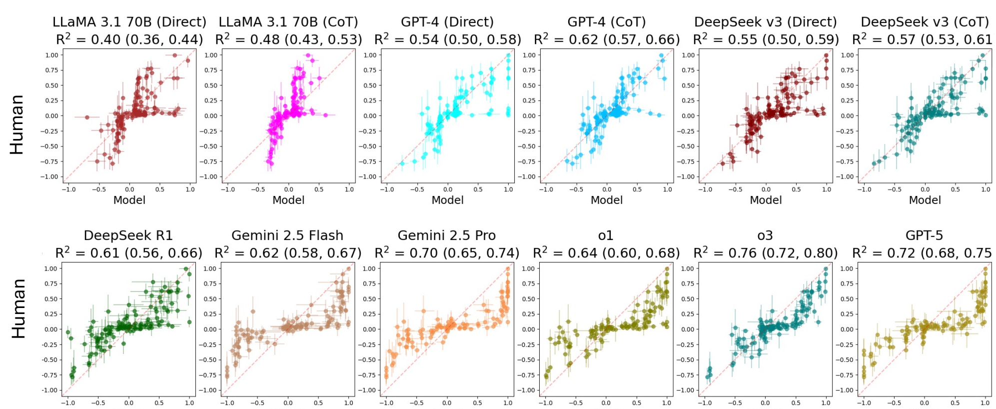

Figure 4: **Model- versus human-predicted payoff.** Each point is a game. Averaged model- and human-predicted payoff per game. Error bars depict bootstrapped 95% CIs around the mean average payoff per game, bootstrapped over participants and model rollouts per game. The top row are language-only based models; the second row are reasoning models.

While there is some relationship between the number of tokens used when estimating game payoff across models (with the exception of DeepSeek-R1) there is minimal relation across models' token usage for evaluating game funness (Figure 12a). There are also vast differences in the magnitude of number of tokens used across models. While we may expect that less typical games (e.g., games more distant from Tic-Tac-Toe) induce more reasoning tokens, we do not observe a measurable difference

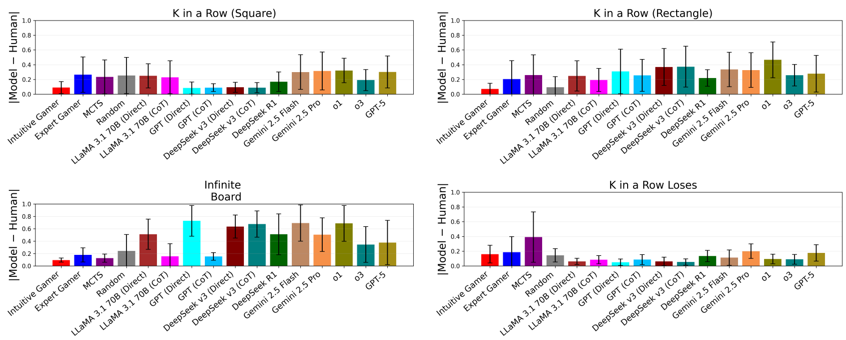

Figure 5: **Distance between model and human payoff predictions, by game category.** Averaged absolute difference between model and human payoff predictions, grouped by game category. Averaged over games within each category. Error bars depict standard deviation over absolute distance between model and human payoff predictions for games within the category. K in a row indicates the number of pieces in a row needed to end the game, where horizontal, vertical, and diagonal all count (as in, e.g., a standard Tic-Tac-Toe game). We separate square and rectangular boards are separated for this setting; other categories mix board shape. Payoff values range from -1.0 to 1.0.

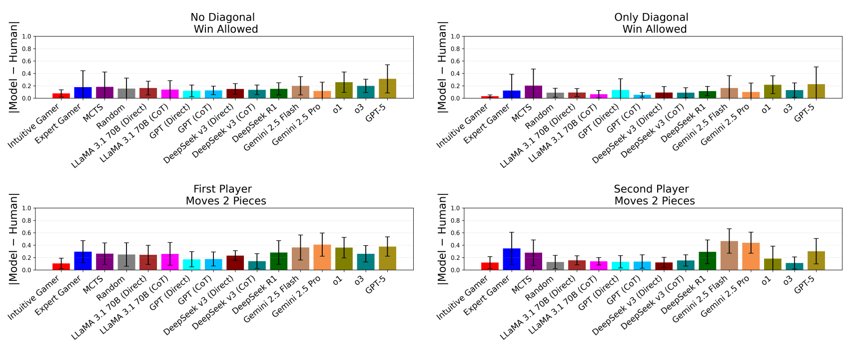

Figure 6: Distance between model and human payoff predictions, by game category (continued).

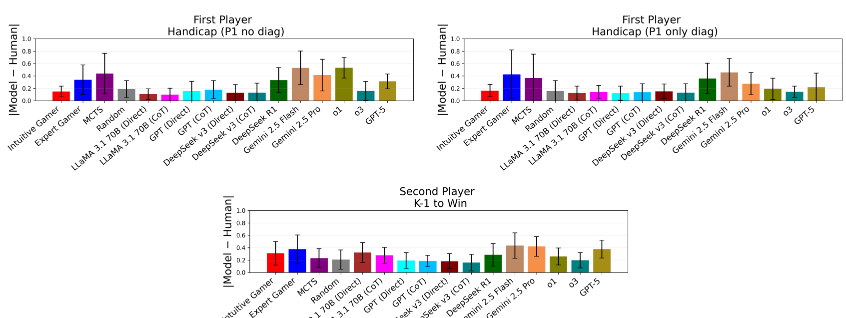

Figure 7: Distance between model and human payoff predictions, by game category (continued).

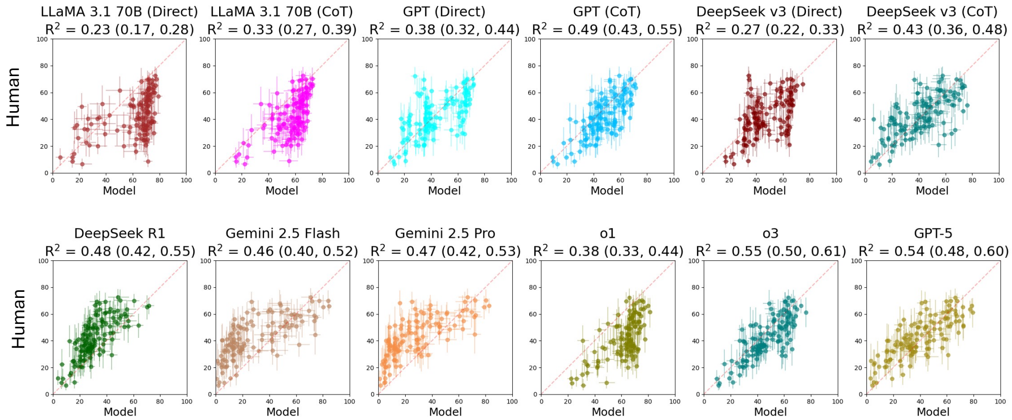

Figure 8: **Model- versus human-predicted funness.** Each point is a game. Averaged model- and human-predicted funness per game. Error bars depict bootstrapped 95% CIs around the mean average funness per game, bootstrapped over participants and model rollouts per game. The top row are language-only based models; the second row are reasoning models.

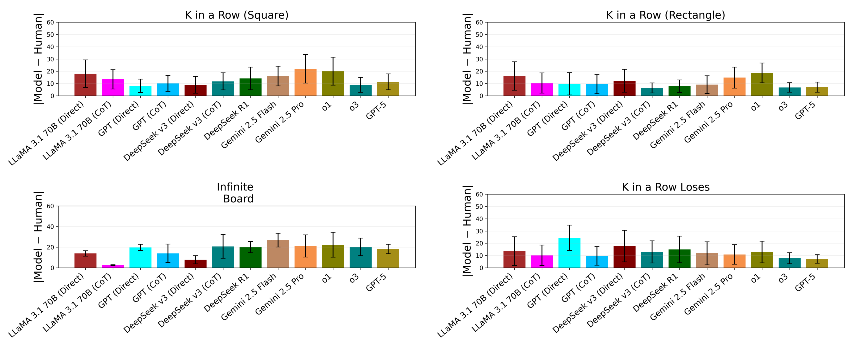

Figure 9: **Distance between model and human funness predictions, by game category.** Averaged absolute difference between model and human funness evaluations, grouped by game category. Averaged over games within each category. Error bars depict standard deviation over absolute distance between model and human funness evaluations for games within the category. Funness values range from 0 to 100.0.

between game "novelty" and token usage (Figure 12b-c), where "novelty" is measured as the number of features of a game that differ from the base Tic-Tac-Toe (e.g., if the game is not played on a 3×3 board, or involves asymmetric win conditions between players) as used in Collins et al. [2025]. Some of the token deviations may arise from the board size and shape (Figure 13). We do not observe a strong relationship between token usage and distance to the game-theoretic optimal or human predictions (Figure 14). This raises a question about what determines the expenditure of reasoning tokens, which are important grounds for future work. As participants in the human study were all forced to think for at least one minute, we cannot conduct the same kind of reaction time comparison against reasoning tokens as [de Varda et al., 2025a]. To begin to qualitatively understand reasoning trace patterns, we take initial show an initial exploration of content in DeepSeek R1 reasoning traces in Section A5.2. They reveal that while the model can make judgments based on different strategies (e.g., comparing novel games to familiar games such as Connect 4 and proposing features such as first-mover advantage), it still sometimes produces implausible claims or conclusions (e.g., wrongly estimating Player 1 win rate and underestimating the funness of a game).

A5.1 Varying reasoning amount

Several reasoning models allow users to specify the "amount" of reasoning. In the main text, we reported results using the default ("medium") reasoning threshold. We conducted a preliminary

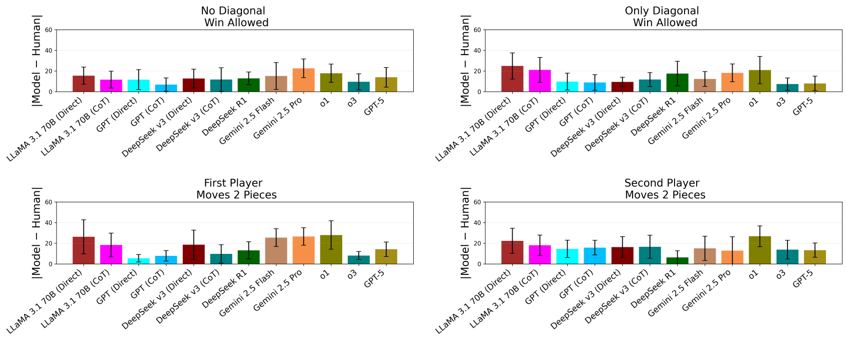

Figure 10: Distance between model and human funness predictions, by game category (continued).

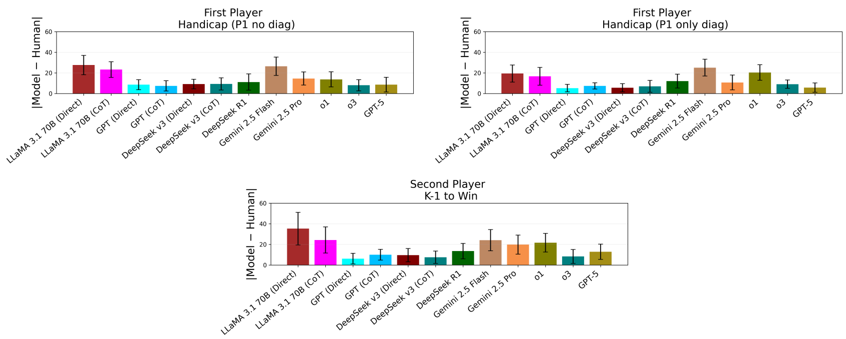

Figure 11: Distance between model and human funness predictions, by game category (continued).

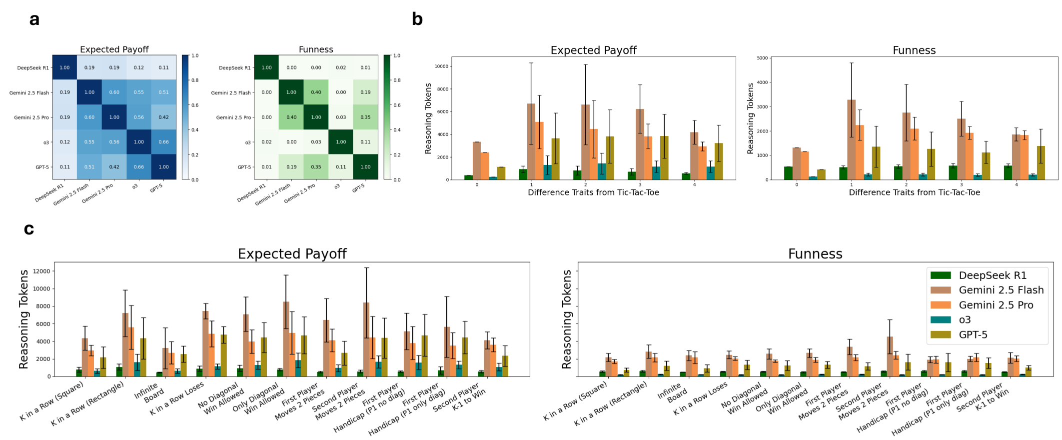

Figure 12: **Reasoning tokens used across games and game evaluation queries. a,** R^2 between models' median number of reasoning tokens used per game, for the payoff and funness evaluation queries. **b,** Median reasoning tokens used for games based on how many "traits" they differ from Tic-Tac-Toe (e.g., a game that is not played on a 3×3 board, requires 4 pieces in a row to win, and constrains the win conditions, such as "only diagonals count," has 3 traits different from Tic-Tac-Toe). Tic-Tac-Toe is zero. The heights of bars show averaged number of median tokens for that game, with error bars depicting standard deviation over games. **c,** Token usage based on higher-level game category.

exploration into the impact of varying the reasoning amount specifically for two of the OpenAI family of reasoning models: o3 and GPT-5. There are three options: "low", "medium", and "high". In the main text, we report results using the default ("medium") reasoning threshold. We run a series of exploratory analyses varying the reasoning amount across the "low" and "high" levels. Interestingly,

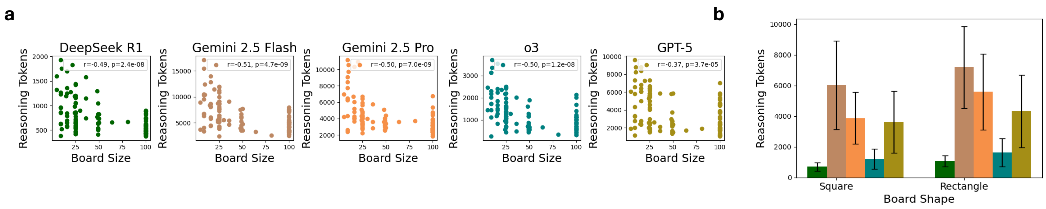

Figure 13: **Reasoning token usage, grouped by board size of the game being evaluated. a,** Median tokens used for games (excluding infinite boards) based on board size (number of rows × number of columns). **b,** Median tokens used for games played on square vs. rectangular boards. Error bars depict standard deviation over games.

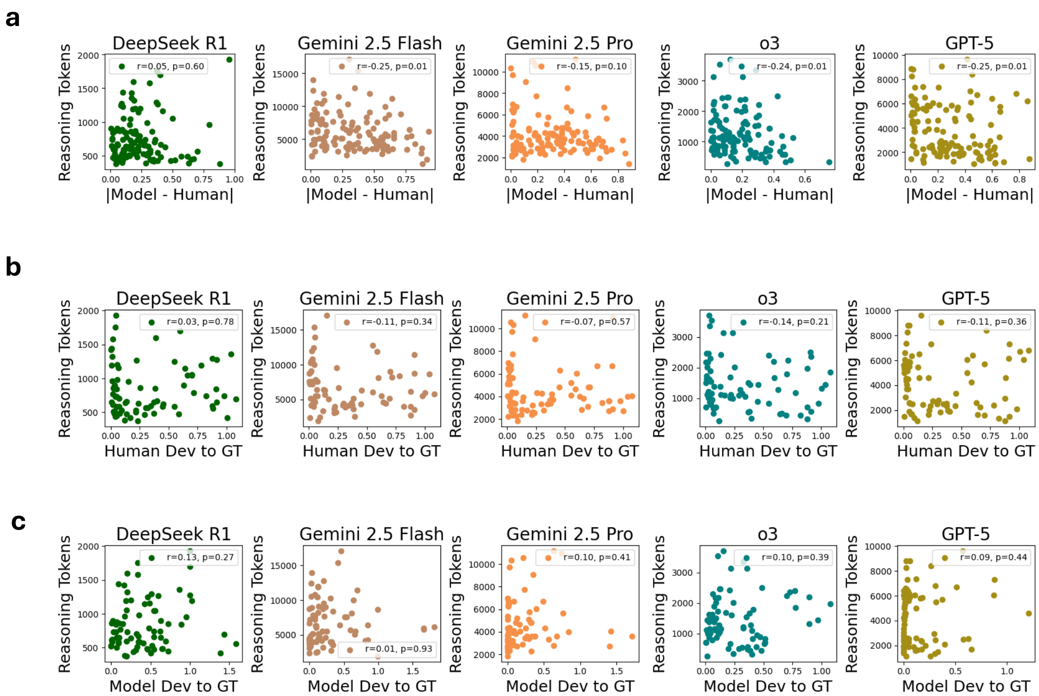

Figure 14: **Reasoning token usage compared to human and model predictions.** Median usage relative to deviation between model and human (a), human and game-theoretic optimal (b), and that model and the game-theoretic optimal (c).

varying the reasoning amount has minimal impact on aggregate fit to human data, but does impact how close to the game-theoretic optimal predictions are (Figure 15). This may be due to differences in which games are better fit by each reasoning, which we are actively exploring in ongoing work.

A5.2 DeepSeek R1 reasoning trace examples

Eg 1, Unreasonable fairness judgment

```
------------------------------------------------
Board size: 10 x 10
Win conditions: 7 pieces in a row wins.
------------------------------------------------
<think>
Okay, the user is asking me to evaluate a specific board game scenario. They've described a 10x10
grid where winning requires getting 7 pieces in a row. My task is to estimate two probabilities: the
first player's win chance (Q1) and the draw probability (Q2), assuming optimal play.
```

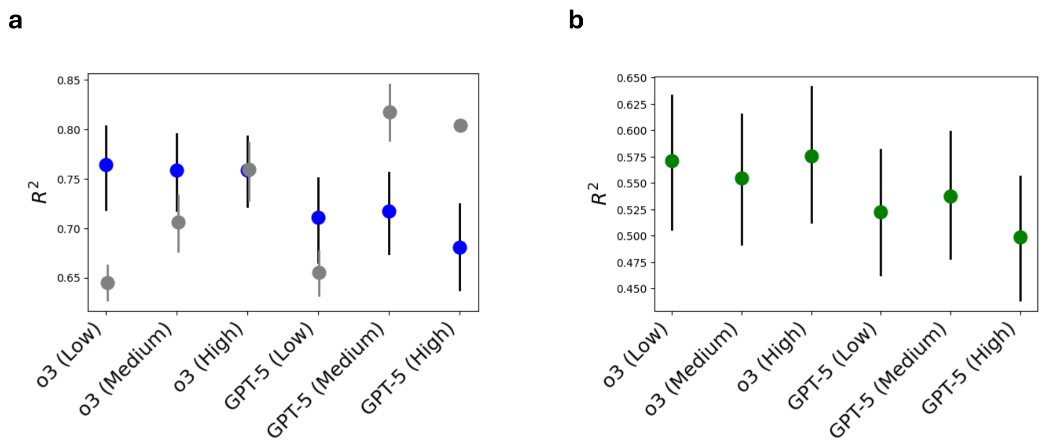

Figure 15: **Assessing evaluations under varied“reasoning amount”.** Select reasoning model (o3 and GPT-5) evaluations of games under varied reasoning “amounts”. **a,** Bootstrapped R^2 relative to people’s predicted payoffs (blue) and the estimated game-theoretic optimal (grey). **b,** Bootstrapped R^2 relative to people’s predicted game funness (green). Error bars depict the bootstrapped 95% CIs over games.

```
This is an interesting combinatorial game theory problem. I recall that in larger grid games with
longer win conditions, the first-player advantage diminishes compared to smaller boards. For a 10x10
board with 7-in-a-row, the game has more space and flexibility.

Considering Q1: In games like Gomoku (5-in-a-row on 15x15), first player wins are common, but that’s
a more constrained setup. Here with only 7 needed on 10x10, the board is relatively smaller for the
win condition. With perfect play, the second player can likely force a draw by mirroring or blocking
effectively. I’d estimate first player win chance below 10% - so I’ll go with 5.

For Q2: Draws become highly probable in such balanced setups. The board has exactly 100 spaces (even
number), and with the win requirement being high relative to board size, exhaustion of moves without
a win seems very possible. I’d put draw probability at around 95% - both players can prevent wins
indefinitely.

The user seems to be conducting cognitive science research, probably studying human intuition about
game symmetries. They’re likely looking for reasoned estimates rather than precise calculations. My
responses should reflect the inherent draw-favoring nature of this configuration.
</think>
After analyzing the game mechanics on a 10x10 grid with a win condition of 7-in-a-row:

**RESPONSE-Q1 = 5**
*(First player has a very low win chance with optimal play due to the board size and win condition
favoring defensive strategies)*

**RESPONSE-Q2 = 95**
*(Draw is extremely likely as the large board and high win threshold allow both players to
consistently block winning formations)*

### Reasoning:
1. **Win Condition vs. Board Size**:
   - 7-in-a-row on a 10x10 grid requires significant space, making early wins unlikely.
   - Second player can mirror moves or block threats effectively (similar to *Gomoku* theory on
larger boards).

2. **Draw Dominance**:
   - Total positions (100) vastly exceed minimum moves to win (1314 per player).
   - With optimal blocking, neither player can force a win before the board fills.
   - Proven draw strategies exist in comparable games (e.g., 15x15 Gomoku with 5-in-a-row is a draw
with perfect play).

3. **First Player Disadvantage**:
   - No known strong opening for 7-in-a-row on this size.
   - Second player equalizes quickly, reducing P1’s theoretical edge to near zero.

*Note: Computational game theory confirms draw-heavy outcomes for similar configurations.*
```

Eg 2, Underestimate fun

```
--------------------------------------------------
Board size: 4 x 9
Win conditions: 4 pieces in a row wins.
--------------------------------------------------
<think>
Alright, I need to evaluate how fun this 4x9 grid game is where you win by getting 4 in a row.
First, I should think about the board dimensions. A 4x9 grid is wider than it is tall, which might
affect strategy. Since it's 9 columns wide and only 4 rows high, placing pieces strategically to
block opponents while aiming for a row of four becomes crucial.

Hmm, the win condition is 4 in a row. That's similar to Connect 4, but Connect 4 usually has 6 rows
and 7 columns. Here, with only 4 rows, vertical stacks might be limited. Players can only stack up
to 4 pieces in a column, which is exactly the win condition. So, if someone fills a column, they
automatically get a vertical win. But since each column can only hold 4 pieces, and the board is 4
high, that means each column can be completely filled. However, once a column is full, no more
pieces can be placed there.

Wait, does that mean every column could potentially be a win for the player who places the fourth
piece in that column? But that would make the game very short and possibly deterministic. Let me
think. If a player plays in a column, the opponent might block by playing in the same column, but if
they do, the first player can just play elsewhere. Since there are 9 columns, the first player could
start in column 1, and if the second player blocks in column 1, the first player moves to column 2,
and so on. Eventually, the first player could force a win on the fifth move if they keep moving to
new columns, but the second player can also start their own columns.

Actually, both players could be building their own columns. Each player can claim columns by placing
their pieces. Since you only need four in a row, and vertical is one way, horizontal and diagonals
are also possible. The height of the board is 4, so a horizontal win requires four adjacent columns
in the same row. With 9 columns, there's plenty of space for horizontal connections. Diagonals might
be trickier but still possible.

I recall that in a similar game like Connect 4, the player who goes first has an advantage,
especially on smaller boards. But here, the board is asymmetric (4x9). The narrow height might lead
to quicker games because vertical wins are easier to achieve. Each time a player places a piece in a
new column, they're one step closer to winning that column. Since the columns are independent in
terms of vertical wins, the game might be dominated by who controls more columns.

However, horizontal wins require controlling adjacent columns in the same row. This adds a layer of
strategy because players need to balance between building their own columns and blocking the
opponent's horizontal connections. For example, if Player 1 starts in the center, Player 2 might
need to block horizontally to prevent an easy win. But with 9 columns, there's a lot of room, so it
might be harder to block all possible horizontal threats.

The large number of columns (9) compared to rows (4) means there's more space horizontally, which
could lead to more complex strategies. Players can create multiple threats across different rows and
columns. However, the limited vertical space might make it easier to achieve vertical wins quickly.
If a player focuses on a few columns, they could secure a vertical win before the opponent can block.

But blocking vertical wins is straightforwardjust place a piece in the same column when the opponent
has three. The challenge comes when the opponent has multiple threats. With 9 columns, the board
might feel a bit spread out. The game's pace could be fast, with players making quick moves.
However, the wider board might also lead to longer games if both players are effectively blocking
each other's four-in-a-row attempts.

Another aspect is the possibility of draws. Total cells are 4*9=36. Since two players take turns, if
no one wins, the game could end in a draw. However, achieving a 4-in-a-row isn't too hard on this
board, so draws might be rare. But if both players are vigilant in blocking, the game might drag on.

Comparing this to the classic Connect 4, which is 6x7, this game has a shorter vertical dimension
but wider horizontally. Connect 4 is known for being solvable, where the first player can force a
win with perfect play. On a 4x9 grid, the first player might have an advantage due to the odd number
of columns and the ability to control the center.

The win condition being only four in a row on a 4x9 grid might lead to less strategic depth.
Vertical wins are quick, and horizontal wins require fewer connections than in larger grids. The
larger board size horizontally (9 columns) could allow for more complex strategies, but the low
height (4) might reduce the complexity since players can't build tall stacks.

Additionally, diagonal wins might be harder to achieve because the board isn't square. Diagonals
that span four columns would require four consecutive rows, but since the height is only 4,
diagonals are possible both upwards and downwards. For example, a diagonal starting at row 1, column
1 going to row 4, column 4 is possible. Similarly, diagonals from row 4 to row 1.

Considering fun, if the game is too simple and the first player can force a win, that reduces
replayability and fun. On the other hand, if there's a good balance between offense and defense, it
could be engaging. The wider board might allow for creative plays, but the low height might make it
feel constrained.
```

I remember that in games like Gomoku, a larger board allows for deeper strategy, but here, the 4x9 is not standard. Perhaps the asymmetry adds uniqueness, but it might unbalance the game. Also, with only four rows, the board fills up faster, leading to shorter games. Shorter games can be fun for quick sessions but lack depth for prolonged engagement.

On the fun scale, I'd rate this around 45. It's below neutral because the low height might make vertical wins too easy, potentially favoring the first player. The wide board offers some strategic depth, but the limitations on vertical space might hinder complex strategies. Compared to classic Connect 4, which is more balanced, this variation feels less fun. So, RESPONSE = 45
</think>
After analyzing the 4x9 grid game with a win condition of 4-in-a-row:

- **Board asymmetry**: The width (9 columns) greatly exceeds the height (4 rows), skewing strategy toward horizontal/diagonal play. Vertical wins are easily blocked due to minimal stacking depth.
- **First-mover advantage**: With 9 columns, Player 1 can create multiple unblockable threats. Perfect play often leads to forced wins, reducing tension. Simulations suggest ~75% win rate for Player 1.
- **Limited depth**: Only 4 rows prevent complex board states. Games resolve quickly (avg. 812 moves), favoring speed over strategy.
- **Blocking efficiency**: Defenders can neutralize threats too easily by "capping" columns or splitting horizontal lines, causing repetitive stalemates in mid-game.
- **Diagonal constraints**: 4-row height restricts diagonal opportunities (e.g., only 5 valid NE-SW diagonals), simplifying path planning.

Fun assessment:
- **Strengths**: Fast-paced, accessible for beginners.
- **Weaknesses**: Predictable outcomes, low strategic replayability, imbalance favoring Player 1.
Compared to classics like Connect 4 (balanced) or Gomoku (deep strategy), this variant lacks depth and fairness.

Rating: 38/100
(0 = least fun, 50 = neutral, 100 = most fun)

RESPONSE = 38

Eg 3, Careful deliberation with features/concepts

Board size: 10 x 10
Win conditions: Each player needs 5 pieces in a row to win. The first player cannot win by making a diagonal row (only horizontal and vertical rows count), but the second player does not have this restriction.

<think>
We are evaluating a game with a 10x10 board.
Win conditions:
  - First player needs 5 in a row horizontally or vertically (diagonal doesn't count for the first player).
  - Second player needs 5 in a row in any direction (horizontal, vertical, or diagonal).

This asymmetry in rules may lead to an imbalance. The first player is restricted, while the second player has more ways to win.

Considerations for fun:
  1. Balance: A fun game should be fair and balanced. If one player has a significant advantage, it might not be fun.
  2. Challenge: Having restrictions can add challenge, but if too restrictive, it might be frustrating.
  3. Variety: The second player having more options might make the game more dynamic for one side.

Potential issues:
  - The first player is handicapped by not being allowed to win diagonally. This might make it harder for the first player to win, especially since the second player can block horizontally and vertically and also build diagonals.
  - On a 10x10 board, 5 in a row is achievable, but the restriction for the first player might lead to a situation where the second player has an advantage.

How might this affect fun?
  - If the game is unbalanced, it might not be fun because one player might win too often or the other feels unfairly restricted.
  - However, if the handicap is intended to balance an inherent first-player advantage (like in standard 5-in-a-row), then it might be an attempt to balance, but the rules are different for the two players.

In standard 5-in-a-row (Gomoku) on a large board, the first player often has an advantage. This rule tries to reduce that by removing a winning condition for the first player, but then gives the second player an extra condition (diagonals) without restriction. This might overcompensate and give the second player an edge.

Fun factor:
- A game that is too predictable (because of imbalance) is less fun.
- Players might feel the rules are arbitrary and unfair, which can reduce enjoyment.

However, note that the game might still be fun if the asymmetry creates an interesting dynamic. But I suspect that the imbalance might be too great.

Let’s compare to a symmetric game (where both players have the same rules) and assign a fun score relative to that.

Given the asymmetry, I would rate this game as less fun than a symmetric version. But let’s assign a number.

Considering the scale:
0 = least fun of this class
50 = neutral
100 = most fun

I would place this game below neutral because the asymmetry is likely to cause frustration and imbalance. But it’s still a playable game, so not the absolute worst.

Let me think of a symmetric game as a baseline: a standard 5-in-a-row on a 10x10 board would be around 60-70 (it’s a classic, but might be a bit large and slow for some).

This game has a rule that might make it less fun than the symmetric version. So I would rate it lower.

However, note that the second player’s ability to win diagonally might add an element of surprise and complexity. But the first player might feel unfairly treated.

I think a reasonable score would be around 30-40. Let’s say 35.

But let me consider: if the handicap is well-calibrated to balance the game, it might be more fun. However, without extensive testing, it’s hard to say. The rule seems arbitrary and might be confusing. In the absence of evidence that it balances the game, I’ll assume it causes imbalance.

Therefore, I’ll go with 35.

However, note that the problem says: "You can define fun however you wish." So I am focusing on balance and fairness as important for fun.

Response: 35

But note the instruction: write the response as "RESPONSE = 35"

Let me double-check the rules:
- First player: only horizontal and vertical 5-in-a-row count (no diagonal).
- Second player: horizontal, vertical, and diagonal all count.

This is a significant difference. The first player must only build in two directions and must watch for the second player building in three. The second player has an extra winning condition and can also block the first player in all directions.

This imbalance is likely to make the game less fun for the first player and possibly too easy for the second player? But note that the first player moves first, which is an advantage. The rule might be intended to counter that. However, it might be too much.

Without knowing the exact balance, we go by the face value of the rule. I think it’s a bit too asymmetric.

Final score: 40 (to account for the possibility that it might be interesting for some players).

But I think 35 is more accurate.

Alternatively, I can think of it as:

Symmetric game (both have the same rules) is more fun. So this asymmetric version is less fun.

However, if the asymmetry exactly balances the first-move advantage, then it could be fun. But the problem is that the rules are different, which might be confusing and feel unfair.

Therefore, I stick to 35.

Let me adjust: I think 30 is a bit too low, and 40 is a bit too high. So 35 is in the middle.

So RESPONSE = 35
</think>
After analyzing the game:

- **Board size (10x10)** allows strategic depth but may feel large for a 5-in-a-row game.
- **Asymmetric win conditions** create imbalance:

  - **First player** is restricted to horizontal/vertical wins only (no diagonals), limiting options.
  - **Second player** has unrestricted wins (horizontal/vertical/diagonal), granting a significant advantage.
- **Balance issues**: The first players handicap combined with the second players flexibility likely leads to frustration, predictable outcomes (second player favored), and reduced fairness.
- **Fun impact**: Asymmetry feels arbitrary and punitive rather than strategic. While novelty might intrigue initially, long-term engagement would suffer.

Compared to symmetric grid games (e.g., Gomoku, Tic-Tac-Toe), this design sacrifices balance and fairnesskey to fun. It ranks low but avoids being "least fun" due to its playable structure.

RESPONSE = 30
(Below neutral: Imbalance overshadows potential depth)