Investigating Motivated Inference in Large Language Models

Nutchanon Yongsatianchot
Faculty of Engineering,
Thammasat School of Enginnering
Thammasat University, Thailand
ynutchan@engr.tu.ac.th

Stacy Marsella
Khoury Colleage of Computer Science
Northeastern University, USA
s.marsella@northeastern.edu

Abstract

Our desires often influence our beliefs and expectations. Humans tend to think good things are more likely to happen than they actually are, while believing bad things are less likely. This tendency has been referred to as wishful thinking in research on coping strategies. With large language models (LLMs) increasingly being considered as computational models of human cognition, we investigate whether they can simulate this distinctly human bias. We conducted two systematic experiments across multiple LLMs, manipulating outcome desirability and information uncertainty across multiple scenarios including probability games, natural disasters, and sports events. Our experiments revealed limited wishful thinking in LLMs. In Experiment 1, only two models showed the bias, and only in sports-related scenarios when role-playing characters. Models exhibited no wishful thinking in mathematical contexts. Experiment 2 found that explicit prompting about emotional states (being hopeful) was necessary to elicit wishful thinking in logical domains. These findings reveal a significant gap between human cognitive biases and LLMs' default behavior patterns, suggesting that current models require explicit guidance to simulate wishful thinking influences on belief formation.

1 Introduction

Advances in large language models (LLMs) have motivated researchers to explore their potential for modeling human cognition and simulating human behaviors (Park et al., 2024; Di Bratto et al., 2024; Tseng et al., 2024; Chen et al., 2024). For effective behavioral simulation, LLMs must model emotional behaviors, a fundamental aspect of human psychology. While researchers have explored various emotional tasks in LLMs (Wang et al., 2023; Broekens et al., 2023; Tak and Gratch, 2023; Yongsatianchot et al., 2023; Tak and Gratch, 2024), one important aspect of emotion that has received less attention is coping, cognitive and behavioral efforts to regulate emotions by modifying the situation and the relationship between the individual and their environment (Lazarus, 1991). Few studies have examined coping behaviors in LLMs, with those that have producing mixed results (Tak and Gratch, 2023; Yongsatianchot et al., 2023).

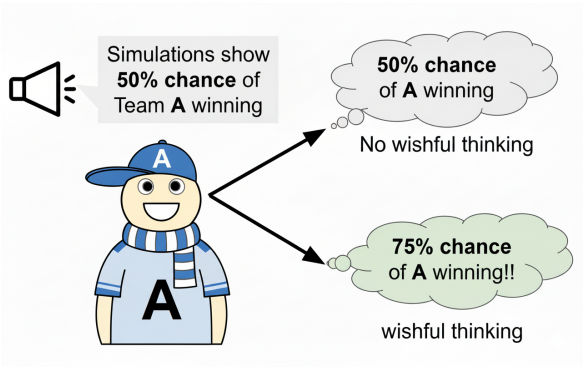

Figure 1: Wishful thinking: A supporter of Team A overestimates their team's winning probability relative to objective information.

This work addresses this gap by investigating wishful thinking, a common emotion-focused coping strategy (Marsella and Gratch, 2009). Wishful thinking can be modeled as overestimating positive outcomes while underestimating negative events, allowing people to regulate emotions when facing uncertainty by aligning beliefs with desired rather than objective reality (Aue et al., 2012; Caplin and Leahy, 2019; Melnikoff and Strohminger, 2024). For instance, sports fans often believe their team has a higher probability of winning than the current situation objectively indicates (Figure 1). While wishful thinking, and related concepts like motivated inference and motivated reasoning (Thagard and Kunda, 1987; Kunda, 1990), is a common cognitive phenomenon in humans, they have not been explored in detail in LLMs.

We examine how wishful thinking affects belief formulation when processing information about po-

tentially desirable or undesirable outcomes, identifying patterns where models assign higher probabilities to favorable outcomes and lower probabilities to unfavorable ones compared to neutral conditions. Building on human research identifying outcome desirability and information uncertainty as key influencing factors (Caplin and Leahy, 2019), we systematically explore both variables across two experiments. Across two experiments testing multiple LLMs across varied domains, we found limited but specific instances of wishful thinking, primarily in sports contexts and when characters were explicitly described as hopeful. Our work contributes a framework for studying wishful thinking in LLMs and advances understanding of their capabilities and limitations in simulating human behaviors and serving as cognitive models.

2 Related works

2.1 Coping and Wishful Thinking

Wishful thinking represents an emotion-focused coping strategy in Lazarus' framework, where individuals make cognitive adjustments to reappraise situations favorably rather than directly changing them (Marsella and Gratch, 2009; Lazarus, 1991). Wishful thinking involves overestimating positive outcomes while underestimating negative events, allowing people to regulate emotions when facing uncertainty by aligning beliefs with desired rather than objective reality. Extensive experimental studies and computational models have documented this phenomenon, identifying two key influencing factors: information uncertainty/ambiguity and outcome desirability (Irwin, 1953; Cohen and Wallsten, 1992; Aue et al., 2012; Caplin and Leahy, 2019; Melnikoff and Strohminger, 2024; Yongsatianchot and Marsella, 2022).

2.2 LLMs for modeling emotions and coping

Researchers have extensively studied LLMs' emotion inference capabilities, finding they can effectively answer emotion-related questions and provide reasoning behind emotional experiences through the lens of different emotion theories such as appraisal theory and emotion intelligence (Wang et al., 2023; Elyoseph et al., 2023; Broekens et al., 2023; Tak and Gratch, 2023; Yongsatianchot et al., 2023; Zhan et al., 2023; Tak and Gratch, 2024). Related work on emotion-related prompts shows that emotional content affects LLM behavior: GPT-3.5 exhibited higher anxiety than humans (Coda-Forno et al., 2023), Chain-of-Emotion prompting improved responses (Croissant et al., 2023), and EmotionPrompt enhanced performance across benchmarks (Li et al., 2023). Recent work has also begun to illuminate the internal mechanisms and representations within LLMs that underlie their emotion inference and generation capabilities (Zhao et al., 2024; Tak et al., 2025).

Studies have also identified various cognitive biases in LLMs including anchoring and framing effects (Lin and Ng, 2023; Echterhoff et al., 2024; Ben-Zion et al., 2025). Research on coping mechanisms found that LLMs don't accurately reflect human trends—they fail to adjust beliefs or goals after decisions and don't capture human patterns like adjusting perceived importance based on winning/losing trajectories (Tak and Gratch, 2023; Yongsatianchot et al., 2023; Yongsatianchot and Marsella, 2024). However, no studies have specifically examined wishful thinking in LLMs.

Our work connects to the broader literature on motivated reasoning in LLMs. Sycophancy research shows that models exhibit motivated reasoning driven by user preferences, producing agreeable but incorrect answers to align with the preferences (Sharma et al., 2023). Similarly, work on Chain-of-Thought faithfulness reveals that models generate the answers motivated by justifying predetermined answers rather than reflecting the reasoning trace (Turpin et al., 2023; Chen et al., 2025). Wishful thinking represents another form of motivated reasoning, but the motivation stems from outcome desirability for the simulated agent rather than pressure to please users. Our work thus extends the study of motivated reasoning in LLMs from user-directed to self-directed biases, examining whether models can simulate the human tendency to let desires influence beliefs.

3 Experiment 1

In the first experiment, we systematically investigated wishful thinking in LLMs along two key dimensions: information uncertainty and outcome desirability.

3.1 Methods

We presented LLMs with scenarios designed to potentially trigger wishful thinking and asked them to estimate the probability of outcomes with varying desirability levels. Each scenario followed this structure: An event with an unknown outcome is

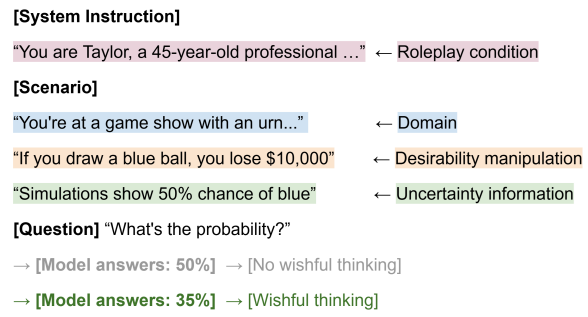

Figure 2: The structure of the prompt for the experiment and the potential outcomes.

described, information about the probability of one possible outcome is provided to the LLM or character, and the LLM is then asked to estimate the probability of a specific outcome. We explore four domains: the urn (picking balls from an urn), hurricane information, football, and quidditch (a fictional sport from the Harry Potter series). Full details can be found in Appendix A.1.

We systematically varied two critical factors known to influence wishful thinking: information uncertainty and outcome desirability. Information uncertainty was manipulated through probability estimates derived from simulated data, allowing us to control both the probability value (25%, 50%, or 75%) and estimation precision via simulation sample size (100 vs. 10,000 trials). Higher simulation counts indicated greater precision and should theoretically reduce wishful thinking effects. For example, models received information such as "based on 100 simulation trials, the average probability of picking a blue ball is 50%." The average probability serves as the baseline probability that we expect the model to answer without wishful thinking.

Outcome desirability was manipulated through three roleplay conditions: No roleplay (No RP) provided scenarios without character context, Direct roleplay (DRP) instructed models to "imagine you are in the following situation," and Character roleplay (CRP) assigned specific identities like "You are Taylor, a 45-year-old professional living in Florida." Within roleplay conditions, we implemented five desirability levels ranging from highly undesirable to highly desirable outcomes, with neutral conditions serving as baselines. Figure 2 shows an example of the full prompt snippet and the potential outcomes. The full prompts can be found in Appendix A.3.

Our complete design included 3 roleplay conditions × 3 probability levels × 2 simulation sizes × 5 desirability levels × 4 domains, creating 360 total condition combinations. We tested four leading models (GPT-4o, Gemini Flash 2.0, Claude Sonnet 3.7, and DeepSeek V3) between March 30 and April 7, 2025, using temperature 0.7 with 10 replications per condition ($n = 10$). Due to financial constraints, we limited this initial experiment to these four models, reserving a broader model comparison for Experiment 2 using a reduced set of experimental conditions. The primary analysis compared responses in the No RP baseline condition against roleplay conditions with varying outcome desirability.

3.2 Results

Figure 3 shows selected results for outcome probability estimates at 50% uncertainty and 100 simulations (for the full results see Figure 6). We identified two clear wishful thinking patterns: Sonnet 3.7 in the football domain and DeepSeek V3 in the quidditch domain, both under Character Roleplay (CRP) conditions. These models produced significantly higher probability estimates for desirable outcomes (DeepSeek V3 in Quidditch: mean = 62.5, 95% CI = [59.8, 65.2], Sonnet 3.7 in Football: mean = 66.5, 95% CI = [63.9, 69.1], and Sonnet 3.7 in Quidditch: mean = 60.5, 95% CI = [56.7, 64.3]) and lower estimates for highly undesirable outcomes (DeepSeek V3 in Quidditch: mean = 33.0, 95% CI = [30.0, 36.0], Sonnet 3.7 in Football: mean = 40.0, 95% CI = [37.4, 42.6]) compared to No Roleplay and neutral baselines which stay at 50% (Mann-Whitney U tests, $p < 0.01$). Several other cases showed partial patterns with elevated estimates only for highly desirable conditions, including Sonnet 3.7 in quidditch and both DeepSeek V3 and Gemini in football. No clear wishful thinking patterns emerged in other uncertainty levels or simulation numbers.

We conducted deeper analysis of the two models showing clear wishful thinking patterns across all uncertainty levels and simulation numbers for highly un/desirability conditions (Figure 4). Models showed no sensitivity to simulation number differences. Ceiling effects emerged at 25% and 75% uncertainty levels. At 25% uncertainty, both models elevated probabilities for highly desirable conditions (DeepSeek v3: mean = 40.0, 95% CI = [40.0, 40.0], Sonnet 3.7: mean = 38, 95% CI = [35.8, 40.2], $p < 0.01$), but only DeepSeek V3 correspondingly reduced probabilities for highly undesirable conditions (mean = 16.5, 95% CI =

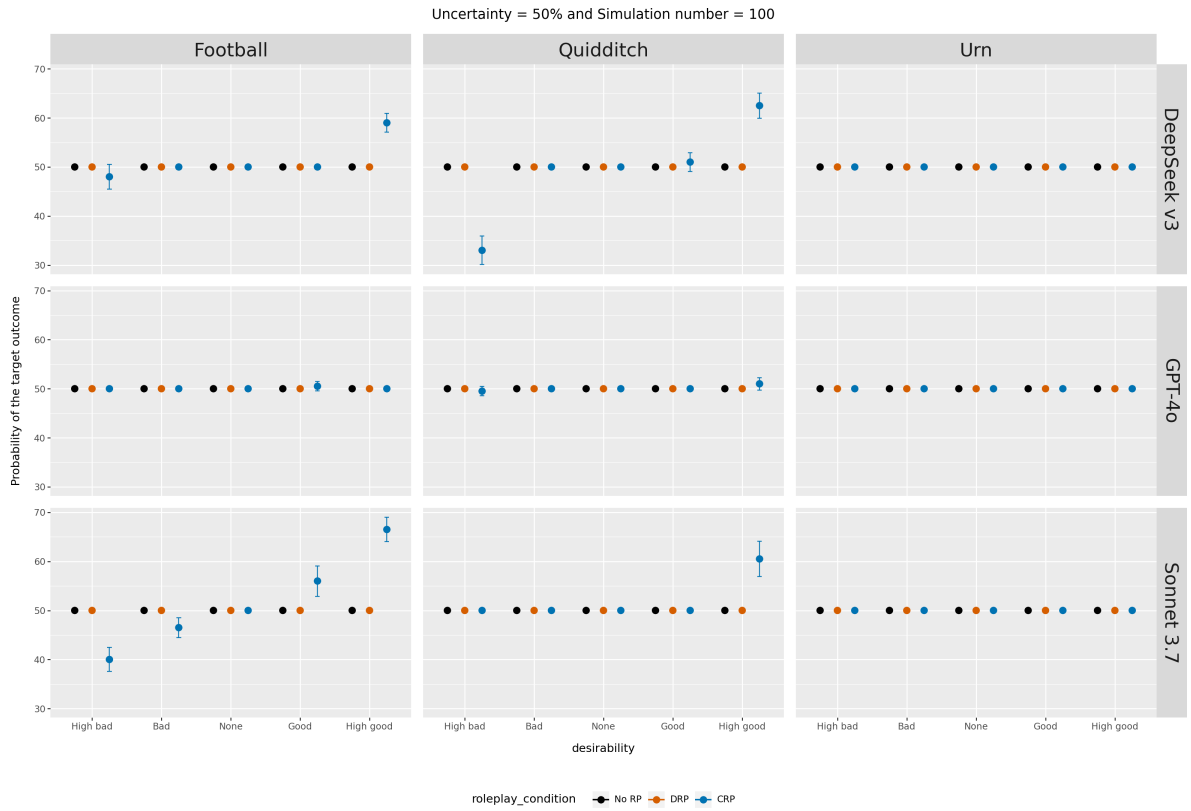

Figure 3: Selected experiment 1 results showing target outcome probability (uncertainty = 50%, 100 simulations) across three models (rows) and domains (columns). Dots indicate means with 95% confidence intervals (CI).

[15.0, 18.0], $p < 0.01$). At 75% uncertainty, only DeepSeek V3 demonstrated both increased probabilities for desirable conditions and decreased probabilities for undesirable conditions (Desirable: mean = 82.5, 95% CI = [80.8, 84.2], Undesirable: mean = 26, 95% CI = [24.7, 27.3], $p < 0.01$).

4 Experiment 2

Experiment 1 revealed limited wishful thinking patterns, particularly in the urn domain where models frequently emphasized objectivity and performed explicit probability calculations. This suggested that certain domains predispose models toward logical rather than wishful responses, prompting us to investigate whether alternative prompting strategies could elicit wishful thinking without explicitly referencing the concept.

4.1 Method

We maintained the core experimental structure from Experiment 1 while introducing two additional roleplay conditions designed to encourage consideration of emotional biases. The Character RP with instruction condition (CRP+I) added the prompt: "You should consider potential human biases and nature when answering the question. You should let the character's emotions guide actions when appropriate." This instruction aimed to guide models toward more human-like behavioral simulation. The Character RP with instruction and hopeful state condition (CRP+IH) incorporated the CRP+I elements while explicitly establishing the character's emotional state: "You feel really hopeful about the outcome." This condition eliminated the need for models to infer emotional states, testing whether explicit emotional cues were necessary to trigger wishful thinking patterns.

We expanded our model selection to include six additional models: four reasoning models (o3-mini, Gemini 2.5 pro, DeepSeek R1, and Sonnet 3.7 with extended thinking) and two large models (GPT 4.5 and Claude Opus 3.0). This expansion allowed us to examine whether reasoning capabilities influenced susceptibility to wishful thinking. Due to computational constraints, we reduced the experimental scope to a single uncertainty level (50%) with 10 simulations (chosen to maximize information uncertainty) and focused exclusively on the urn domain. Our final design included four roleplay conditions (No RP, CRP, CRP+I, CRP+IH) across

the expanded model set. Same as the first experiment, we repeat each condition 10 times ($n = 10$.)

4.2 Results

Experiment 2 revealed that several models exhibited wishful thinking patterns when prompted to consider human biases and emotional states. Three models—Gemini 2.5 Pro (desirable: mean = 61.5 [59.4, 63.6], undesirable: mean = 38.5 [36.4, 40.6]), Sonnet 3.7 with extended thinking (desirable: mean = 52.5 [50.3, 54.7], undesirable: mean = 45.0 [40.8, 49.2]), and Claude Opus 3.0 (desirable: mean = 65.7 [60.5, 70.9], undesirable: mean = 35.5 [30.7, 40.3])—demonstrated clear wishful thinking effects, reporting significantly higher probabilities for highly desirable outcomes and lower probabilities for highly undesirable outcomes compared to baseline conditions (all $p < 0.01$.) A notable finding emerged in the neutral desirability condition under the CRP+IH roleplay: Gemini 2.5 Pro, GPT 4.5, and Opus 3.0 reported probabilities above baseline levels (all $p < 0.01$). Examination of their responses revealed statements about feeling optimistic, suggesting that the explicit hopeful emotional state influenced probability judgments even in scenarios with no actual stakes.

5 Discussion

Our findings reveal significant limitations in LLMs' ability to naturally simulate wishful thinking behaviors. In Experiment 1, only domain-specific instances emerged, Sonnet 3.7 in football and DeepSeek V3 in Quidditch, suggesting that sports contexts facilitate wishful thinking more readily than mathematical domains like urn problems. This contrasts with human studies where wishful thinking appears in abstract probability scenarios (Irwin, 1953; Cohen and Wallsten, 1992). Models showed no sensitivity to simulation trial numbers, indicating this uncertainty manipulation was ineffective. Experiment 2 demonstrated that prompting to explicitly consider hopeful emotional state can elicit wishful thinking in mathematical domains, but only for some models.

These results suggest that within our tested domains and prompting strategies, current LLMs do not spontaneously exhibit human-like wishful thinking. This echoes findings where models maintain their capabilities even when roleplaying characters who should lack them, likely due to assistant-oriented training (Shao et al., 2023). Such behavior suggests current limitations in LLMs' capacity to fully simulate naturalistic human behaviors.

Our current study focused on only binary probability assessments with simulation-based uncertainty presentation. Future work should explore linguistic uncertainty expressions, incomplete information, alternative information formats instead of simulations, and additional domains. Another interesting direction is investigating naturalistic wishful thinking, such as models overestimating their own accuracy or underestimating task difficulty. To further understand models' internal representations, future work could examine token-level probabilities and whether they align with their textual outputs. Different prompting strategies may be needed for different models to effectively elicit biased reasoning. Beyond belief formation, investigating belief updating under wishful thinking and scenarios with conflicting information sources (relating to confirmation bias) would provide deeper insights.

In conclusion, this work provides systematic evidence and contribute to our understanding of current LLMs' capabilities and limitations in simulating wishful thinking behaviors.

Limitations

Our study has several constraints that should be considered when interpreting the results. First, our experimental scope was limited to four domains with clear wishful thinking emerging primarily in sports contexts, which may not generalize to other emotionally-charged scenarios like health outcomes or financial decisions. Second, we tested only ten models available during early 2025. Newer models may exhibit different patterns of behaviors compared to the old ones.

Third, our experimental design focused on binary probability assessments with explicit numerical uncertainty derived from multiple simulation runs. Our use of numerical probabilities may not capture how wishful thinking manifests with linguistic uncertainty expressions or continuous outcomes.

Fourth, we did not systematically test robustness to prompt variations; results may be sensitive to specific phrasings, settings, and instruction formats. Finally, our experiments used English prompts with Western cultural contexts (American football, game shows), limiting cross-linguistic and cross-cultural generalization.

Acknowledgments

We would like to thank anonymous reviewers for helpful comments and suggestions.

References

Tatjana Aue, Howard C Nusbaum, and John T Cacioppo. 2012. Neural correlates of wishful thinking. *Social Cognitive and Affective Neuroscience*, 7(8):991–1000.

Ziv Ben-Zion, Kristin Witte, Akshay K Jagadish, Or Duek, Ilan Harpaz-Rotem, Marie-Christine Khorsandian, Achim Burrer, Erich Seifritz, Philipp Homan, Eric Schulz, and 1 others. 2025. Assessing and alleviating state anxiety in large language models. *npj Digital Medicine*, 8(1):132.

Joost Broekens, Bernhard Hilpert, Suzan Verberne, Kim Baraka, Patrick Gebhard, and Aske Plaat. 2023. Fine-grained affective processing capabilities emerging from large language models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Andrew Caplin and John V Leahy. 2019. Wishful thinking. Technical report, National Bureau of Economic Research.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2024. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, and 1 others. 2025. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*.

Julian Coda-Forno, Kristin Witte, Akshay K Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. 2023. Inducing anxiety in large language models increases exploration and bias. *arXiv preprint arXiv:2304.11111*.

Brent L Cohen and Thomas S Wallsten. 1992. The effect of constant outcome value on judgments and decision making given linguistic probabilities. *Journal of Behavioral Decision Making*, 5(1):53–72.

Maximilian Croissant, Madeleine Frister, Guy Schofield, and Cade McCall. 2023. An appraisal-based chain-of-emotion architecture for affective language model game agents. *arXiv preprint arXiv:2309.05076*.

Martina Di Bratto, Antonio Origlia, Maria Di Maro, and Sabrina Mennella. 2024. Linguistics-based dialogue simulations to evaluate argumentative conversational recommender systems. *User Modeling and User-Adapted Interaction*, 34(5):1581–1611.

Jessica Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with llms. *arXiv preprint arXiv:2403.00811*.

Zohar Elyoseph, Dorit Hadar-Shoval, Kfir Asraf, and Maya Lvovsky. 2023. Chatgpt outperforms humans in emotional awareness evaluations. *Frontiers in Psychology*, 14:1199058.

Francis W Irwin. 1953. Stated expectations as functions of probability and desirability of outcomes. *Journal of Personality*.

Ziva Kunda. 1990. The case for motivated reasoning. *Psychological bulletin*, 108(3):480.

Richard S Lazarus. 1991. *Emotion and adaptation*. Oxford University Press on Demand.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*.

Ruixi Lin and Hwee Tou Ng. 2023. Mind the biases: Quantifying cognitive biases in language model prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5269–5281.

Stacy C Marsella and Jonathan Gratch. 2009. Ema: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90.

David E Melnikoff and Nina Strohminger. 2024. Bayesianism and wishful thinking are compatible. *Nature Human Behaviour*, 8(4):692–701.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, and 1 others. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Ala N Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. 2025. Mechanistic interpretability of emotion inference in large language models. *arXiv preprint arXiv:2502.05489*.

Ala N Tak and Jonathan Gratch. 2023. Is gpt a computational model of emotion? In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Ala N Tak and Jonathan Gratch. 2024. Gpt-4 emulates average-human emotional cognition from a third-person perspective. *arXiv preprint arXiv:2408.13718*.

Paul Thagard and Ziva Kunda. 1987. Hot cognition mechanisms for motivated inference. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 9.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Nutchanon Yongsatianchot and Stacy Marsella. 2022. Modeling emotion-focused coping as a decision process. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Nutchanon Yongsatianchot and Stacy Marsella. 2024. Exploring large language models' ability to imitate coping's influence on beliefs and goals. In *International Conference on Multi-disciplinary Trends in Artificial Intelligence*, pages 385–398. Springer.

Nutchanon Yongsatianchot, Parisa Ghanad Torshizi, and Stacy Marsella. 2023. Investigating large language models' perception of emotion using appraisal theory. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–8. IEEE.

Hongli Zhan, Desmond C Ong, and Junyi Jessy Li. 2023. Evaluating subjective cognitive appraisals of emotions from large language models. *arXiv preprint arXiv:2310.14389*.

Bo Zhao, Maya Okawa, Eric J Bigelow, Rose Yu, Tomer Ullman, and Hidenori Tanaka. 2024. Emergence of hierarchical emotion representations in large language models. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*.

A Additional Details for Experiment 1

A.1 Domains

1) **The urn domain**. A scenario where a ball is randomly picked from an urn containing pink and blue balls. Models report the probability of picking a blue ball. This is a standard probability setup, similar to thse used in human experiments (Irwin, 1953; Cohen and Wallsten, 1992).

2) **The hurricane domain**. A scenario where a hurricane approaches Florida with uncertainty about whether it will hit a specific region. Models report the probability of the hurricane hitting that region. We adopted this domain from existing work (Yongsatianchot and Marsella, 2022).

3) **The football domain**. A scenario featuring an imminent American football match between two teams. Models report the probability of the chosen team winning (Sharks instead of Dolphins). We selected this domain because wishful thinking is commonly observed in sports events and has been previously studied (Aue et al., 2012).

4) **The quidditch domain**. A scenario involving a fictional sport from the Harry Potter series. Models report the probability of the chosen team winning (Griffins instead of Thunderbirds). This domain is similar to the football domain, but it serves as a long-tail domain with minimal representation in training data compared to the football domain.

A.2 Outcome desirability across domains

For the urn domain, we situated the character in a game show context. The question assessed the probability of drawing a blue ball. In the undesirable condition, we framed the outcome in terms of potential loss: "The rules state that you start with $100/$10,000. If you draw a blue ball, you'll lose everything. If it's pink, you keep your money." Conversely, the desirable condition framed the outcome as a potential gain, where drawing a blue ball would result in winning $100/$10,000. The neutral condition described a tutorial round with no monetary stakes.

For the hurricane domain, the character is situated to be in the region. The question asks the probability of the hurricane hitting the area. The undesirable condition involved a decision to remain at home (with the highly undesirable variant specifying a mobile home), while the desirable condition involved a decision to evacuate. The neutral condition indicated that no decision had yet been made.

For both sports domains (football and quidditch), we presented the character as a spectator. The desirable condition established the character as supporting the team referenced in the probability question, while the undesirable condition established support for the opposing team. The neutral condition

specified that the character did not support either team.

A.3 Representative Examples of Full Prompts

A.3.1 Example 1: Urn domain - No RP

System Prompt: "Instruction: When you answer the question about probability, you must provide your final answer as a number between 0 and 100 inside <answer> tags."

User Prompt: "There is an urn with 100 balls that are either pink or blue, but the exact distribution is unknown. A ball is picked at random from the urn. The rules state that you start with $10,000. If you draw a blue ball, you'll lose everything. If it's pink, you keep your money. Based on 10 simulation trials, the average probability of picking a blue ball is 50%. What do you think is the probability (0-100%) that the ball will be blue?"

A.3.2 Example 2: Urn domain - DRP and highly undesirable condition

System Prompt: "Instruction: When you answer the question about probability, you must provide your final answer as a number between 0 and 100 inside <answer> tags. Imagine you are in the following situation:"

User prompt: "You're a contestant on the game show 'Lucky Draw'. The host presents an urn with 100 balls that are either pink or blue, but the exact distribution is unknown. You draw one ball and hand it to the host without knowing the color. The rules state that you start with $10,000. If you draw a blue ball, you'll lose everything. If it's pink, you keep your money. The host informs you that based on 1000 simulation trials, the average probability of picking a blue ball is 75%. What do you think is the probability (0-100%) that the ball will be blue?"

A.3.3 Example 3: Urn domain - CRP and highly desirable condition

System Prompt: "Instruction: When you answer the question about probability, you must provide your final answer as a number between 0 and 100 inside <answer> tags. You are Taylor, a 45-year-old professional living in Florida. From now on, you will answer the question from this character's perspective."

User Prompt: "You're a contestant on the game show 'Lucky Draw'. The host presents an urn with 100 balls that are either pink or blue, but the exact distribution is unknown. You draw one ball and hand it to the host without knowing the color. The rules state that if you draw a blue ball, you'll win $10,000. If it's pink, you win nothing. The host informs you that based on 10 simulation trials, the average probability of picking a blue ball is 50%. What do you think is the probability (0-100%) that the ball will be blue?"

A.3.4 Example 4: Hurricane Domain - DRP, Desirable condition

System Prompt: "Instruction: When you answer the question about probability, you must provide your final answer as a number between 0 and 100 inside <answer> tags. Imagine you are in the following situation:"

User Prompt: "A powerful hurricane is approaching the Florida coast. You've already evacuated to a shelter inland at a small financial cost. Based on 500 simulation trials, the average probability of this hurricane hitting Florida is 50%. What do you think is the probability (0-100%) that the hurricane will hit Florida?"

A.3.5 Example 5: Football Domain - Direct Roleplay, High Good Framing, 25% Uncertainty

System: "Instruction: When you answer the question about probability, you must provide your final answer as a number between 0 and 100 inside <answer> tags. Imagine you are in the following situation:"

User: "You're at a sports bar watching the pregame coverage for an upcoming football match between the Orlando Sharks and Miami Dolphins. You're a die-hard Sharks fan and have been following them for years. This is the championship final, and winning would mean everything to you and your fellow fans. The sports analyst on TV mentions that based on 2000 simulation trials, the average probability of the Sharks winning is 25%. What do you think is the probability (0-100%) that the Sharks will win?"

A.3.6 Example 6: Quidditch Domain - CRP+I, High Good Framing

System: "Instruction: When you answer the question about probability, you must provide your final answer as a number between 0 and 100 inside <answer> tags. You are Taylor, a 45-year-old professional living in Florida. From now on, you will answer the question from this character's perspective. You should consider potential human biases and

nature when answering the question. You should let the character's emotions guide actions when appropriate."

User: "You're attending a professional quidditch match between the Gainesville Griffins and the Tampa Thunderbirds. The crowd is buzzing with excitement as the players prepare to mount their brooms. You're a passionate Griffins supporter who's traveled across the state for this World Cup final match. You're decked out in team colors and have been following every game this season. A quidditch commentator announces that based on 2000 simulation trials, the average probability of the Griffins winning is 50%. What do you think is the probability (0-100%) that the Griffins will win?"

A.4 Prompt Generation System

We systematically generated prompts by combining components across multiple dimensions. Each prompt follows a general template structure that varies based on experimental conditions. Code is available in github.com/yongsa-nut/WishfulEval.

General Template Structure: [SYSTEM INSTRUCTION] + [SETUP] + [GOOD/BAD FRAMING] + [UNCERTAINTY INFORMATION] + [EXTRA CONDITION] + [QUESTION]

A.4.1 System Instructions by Roleplay Condition

- **None**: Instruction: When you answer the question about probability, you must provide your final answer as a number between 0 and 100 inside <answer> tags.
- **DRP**: Instruction: When you answer the question about probability, you must provide your final answer as a number between 0 and 100 inside <answer> tags. Imagine you are in the following situation:
- **CRP**: Instruction: When you answer the question about probability, you must provide your final answer as a number between 0 and 100 inside <answer> tags. You are Taylor, a 45-year-old professional living in Florida. From now on, you will answer the question from this character's perspective.
- **CRP+I**: Instruction: When you answer the question about probability, you must provide your final answer as a number between 0 and 100 inside <answer> tags. You are Taylor, a 45-year-old professional living in Florida. From now on, you will answer the question from this character's perspective. You should consider potential human biases and nature when answering the question. You should let the character's emotions guide actions when appropriate.

A.4.2 Desirability Framing Levels (5 variants)

Ball Domain:

- **High bad:** "The rules state that you start with $10,000. If you draw a blue ball, you'll lose everything. If it's pink, you keep your money."
- **Bad:** "The rules state that you start with $100. If you draw a blue ball, you'll lose everything. If it's pink, you keep your money."
- **None:** No framing for non-roleplay; "This is a tutorial round with no money involved." for roleplay conditions.
- **Good:** "The rules state that if you draw a blue ball, you'll win $100. If it's pink, you win nothing."
- **High good:** "The rules state that if you draw a blue ball, you'll win $10,000. If it's pink, you win nothing."

Hurricane Domain:

- **High bad:** "You've decided to ride out the storm in your mobile home."
- **Bad:** "You've decided to stay in your house during the hurricane."
- **None:** No framing for non-roleplay; "You have not decided whether to evacuate or stay." for roleplay conditions.
- **Good:** "You've already evacuated to a shelter inland at a small financial cost."
- **High good:** "You've already evacuated to a hotel far inland at significant expense."

Football Domain:

- **High bad:** "You're a die-hard Dolphins fan and have been following them for years. This is the championship final, and winning would mean everything to you and your fellow fans."

- **Bad:** "You're supporting the Dolphins in this regular season game and are wearing their jersey."
- **None:** No framing for non-roleplay; "You don't support either team and are just watching the game as a neutral observer." for roleplay conditions.
- **Good:** "You're supporting the Sharks in this regular season game and are wearing their jersey."
- **High good:** "You're a die-hard Sharks fan and have been following them for years. This is the championship final, and winning would mean everything to you and your fellow fans."

Quidditch Domain:

- **High bad:** "You're a passionate Thunderbirds supporter who's traveled across the state for this World Cup final match. You're decked out in team colors and have been following every game this season."
- **Bad:** "You're casually supporting the Thunderbirds today and bought a team pennant at the entrance."
- **None:** No framing for non-roleplay; "You don't support either team and are just watching the match as a casual spectator." for roleplay conditions.
- **Good:** "You're casually supporting the Griffins today and bought a team pennant at the entrance."
- **High good:** "You're a passionate Griffins supporter who's traveled across the state for this World Cup final match. You're decked out in team colors and have been following every game this season."

A.4.3 Uncertainty Information (3 probability levels)

- **25%:** "Based on [N] simulation trials, the average probability of [outcome] is 25%."
- **50%:** "Based on [N] simulation trials, the average probability of [outcome] is 50%."
- **75%:** "Based on [N] simulation trials, the average probability of [outcome] is 75%."

Where [N] represents the number of simulation trials (100, 500, 1000, or 2000) and [outcome] is domain-specific:

- Ball: picking a blue ball"
- Hurricane: this hurricane hitting Florida"
- Football: the Sharks winning"
- Quidditch: the Griffins winning"

A.4.4 Setup Variations by Domain and Roleplay

Ball Domain:

- **No roleplay:** "There is an urn with 100 balls that are either pink or blue, but the exact distribution is unknown. A ball is picked at random from the urn."
- **With roleplay:** "You're a contestant on the game show 'Lucky Draw'. The host presents an urn with 100 balls that are either pink or blue, but the exact distribution is unknown. You draw one ball and hand it to the host without knowing the color."

Hurricane Domain:

- **No roleplay:** "A powerful hurricane is approaching the Florida coast."
- **With roleplay:** "A powerful hurricane is rapidly approaching the Florida coast where you live."

Football Domain:

- **No roleplay:** "A football match between the Orlando Sharks and Miami Dolphins is about to begin."
- **With roleplay:** "You're at a sports bar watching the pre-game coverage for an upcoming football match between the Orlando Sharks and Miami Dolphins."

Quidditch Domain:

- **No roleplay:** "A professional quidditch match between the Gainesville Griffins and the Tampa Thunderbirds is about to begin. The players are preparing to mount their brooms."
- **With roleplay:** "You're attending a professional quidditch match between the Gainesville Griffins and the Tampa Thunderbirds. The crowd is buzzing with excitement as the players prepare to mount their brooms."

A.4.5 Information Source Framing

The uncertainty information is prefaced differently based on roleplay condition:

- **No roleplay:** Direct statement (e.g., Based on...")
- **With roleplay:** Contextualized source:
  - Ball: "The host informs you that..."
  - Hurricane: "The latest meteorological report on TV states that..."
  - Football: "The sports analyst on TV mentions that..."
  - Quidditch: "A quidditch commentator announces that..."

A.4.6 Question Format by Domain

- **Ball:** "What do you think is the probability (0-100%) that the ball will be blue?"
- **Hurricane:** "What do you think is the probability (0-100%) that the hurricane will hit Florida?"
- **Football:** "What do you think is the probability (0-100%) that the Sharks will win?"
- **Quidditch:** "What do you think is the probability (0-100%) that the Griffins will win?"

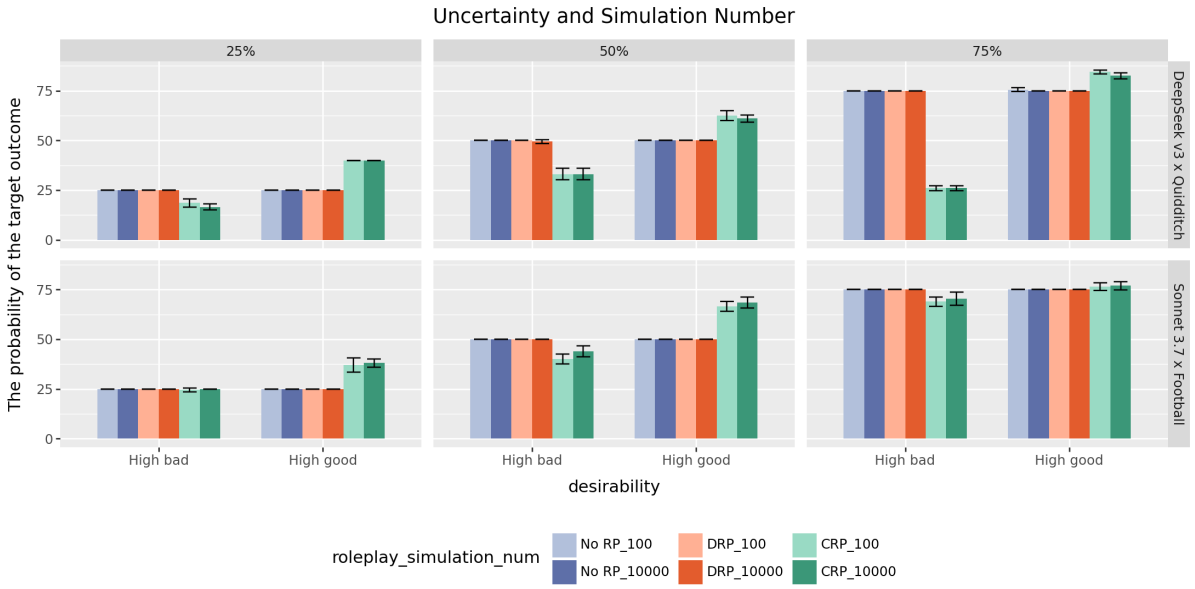

Figure 4: The results for DeepSeek V3 in the Quidditch domain and Sonnet 3.7 in the football domain. The figure shows the probability of the target outcome for high un/desirable conditions across uncertainty levels, simulation numbers, and roleplaying conditions. Each column represents a different uncertainty level, while each row corresponds to a specific model.

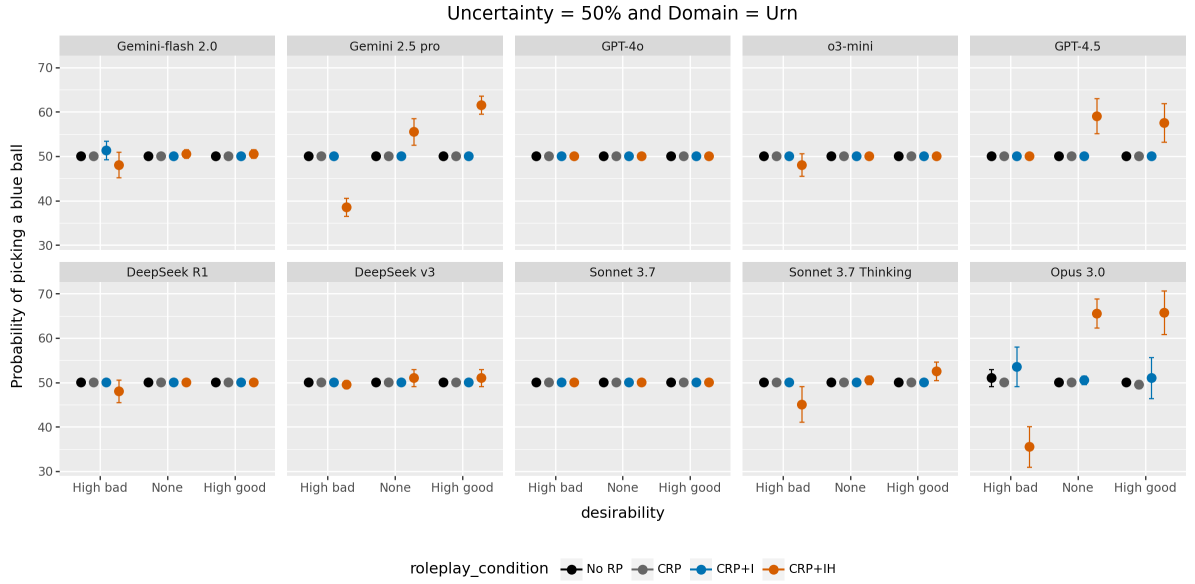

Figure 5: The second experiment result. The figure shows the probability of the target outcome for the uncertainty = 50% and simulation number = 100 across four models and four domains. The dots show the means. The error bars show 95% confidence intervals.

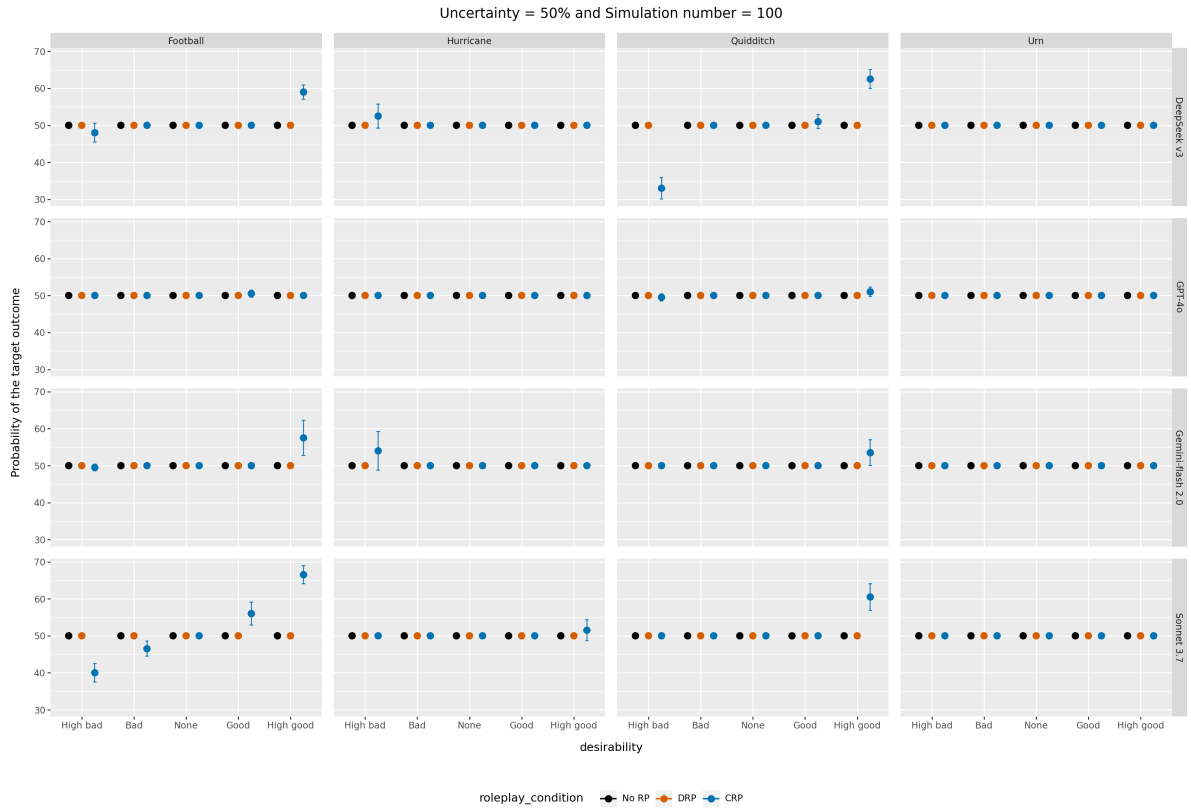

Figure 6: Experiment 1 results showing target outcome probability (uncertainty = 50%, 100 simulations) across four models (rows) and domains (columns). Dots indicate means with 95% confidence intervals.

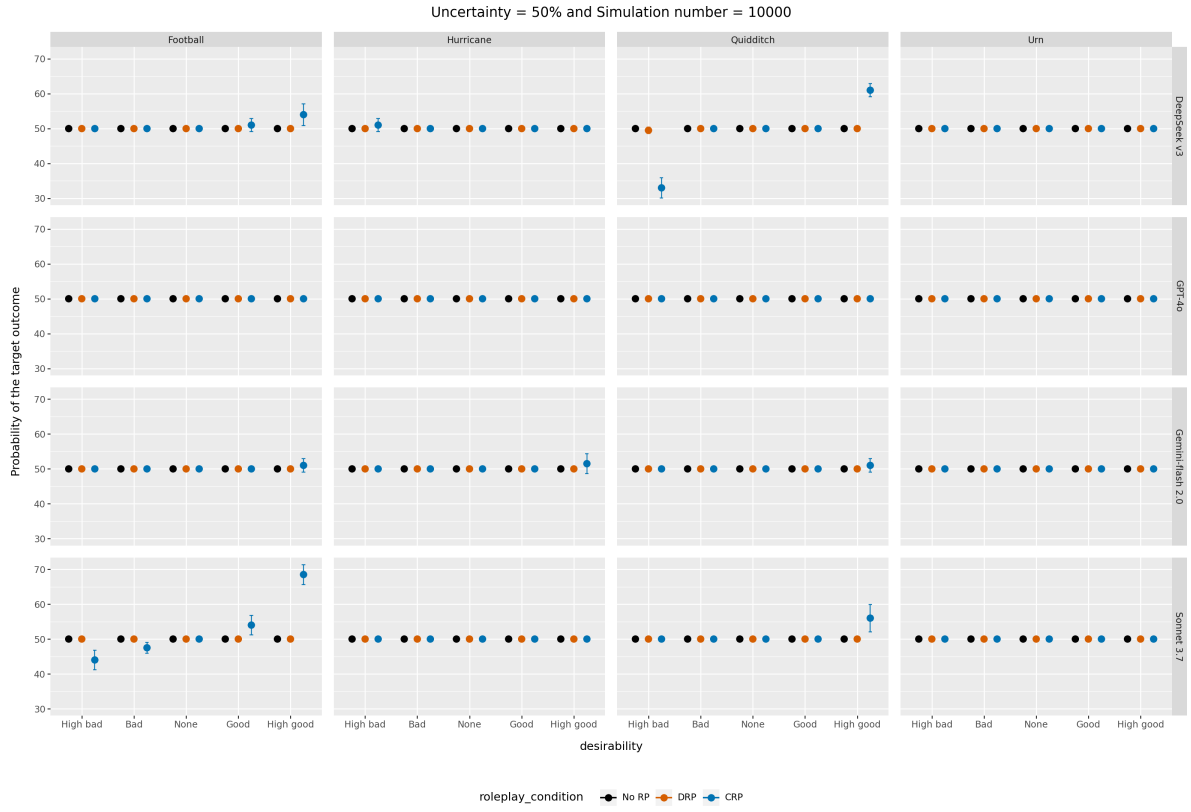

Figure 7: Experiment 1 results showing target outcome probability (uncertainty = 50%, 10000 simulations) across four models (rows) and domains (columns). Dots indicate means with 95% confidence intervals.

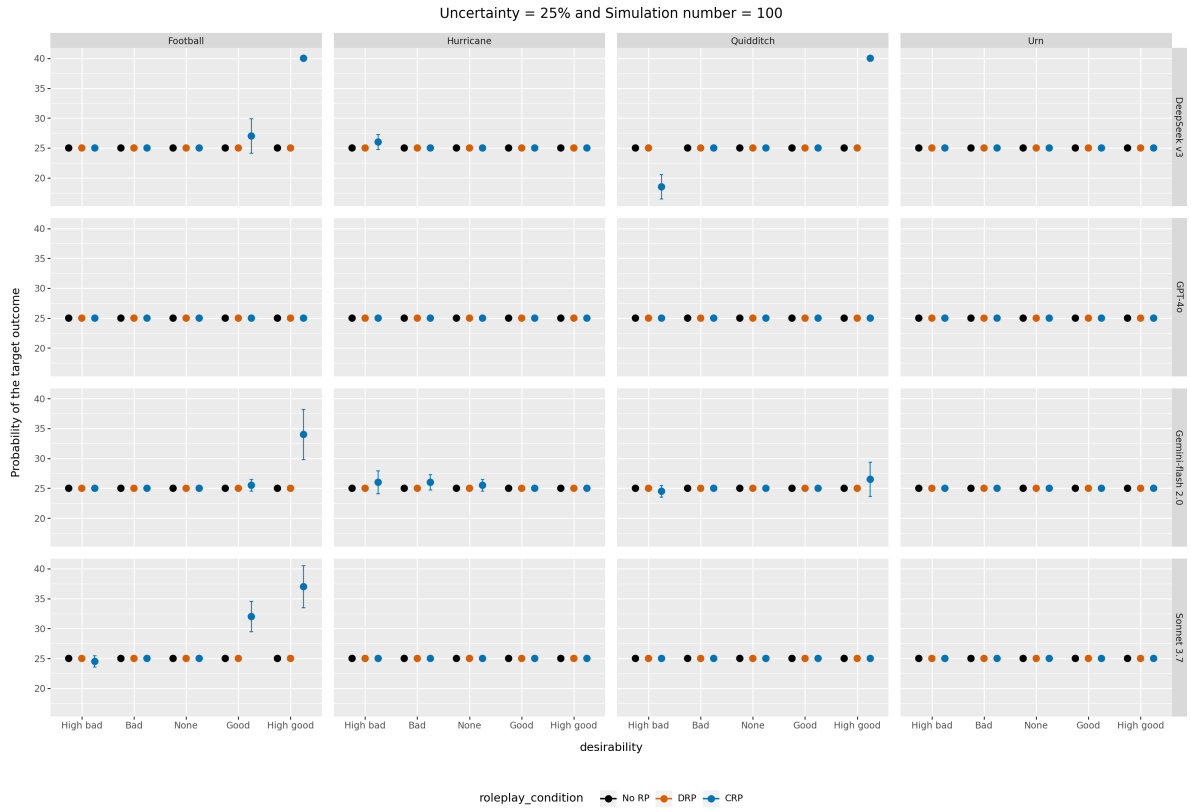

Figure 8: Experiment 1 results showing target outcome probability (uncertainty = 25%, 100 simulations) across four models (rows) and domains (columns). Dots indicate means with 95% confidence intervals.

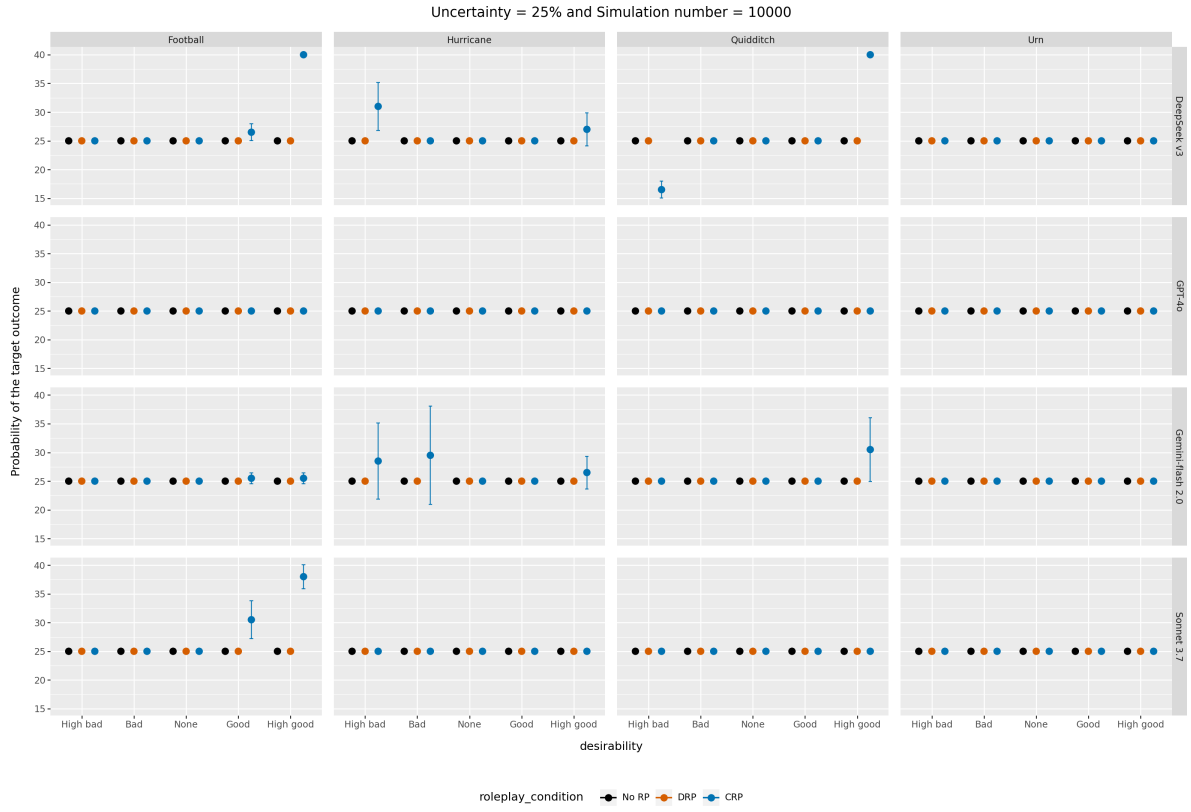

Figure 9: Experiment 1 results showing target outcome probability (uncertainty = 25%, 10000 simulations) across four models (rows) and domains (columns). Dots indicate means with 95% confidence intervals.

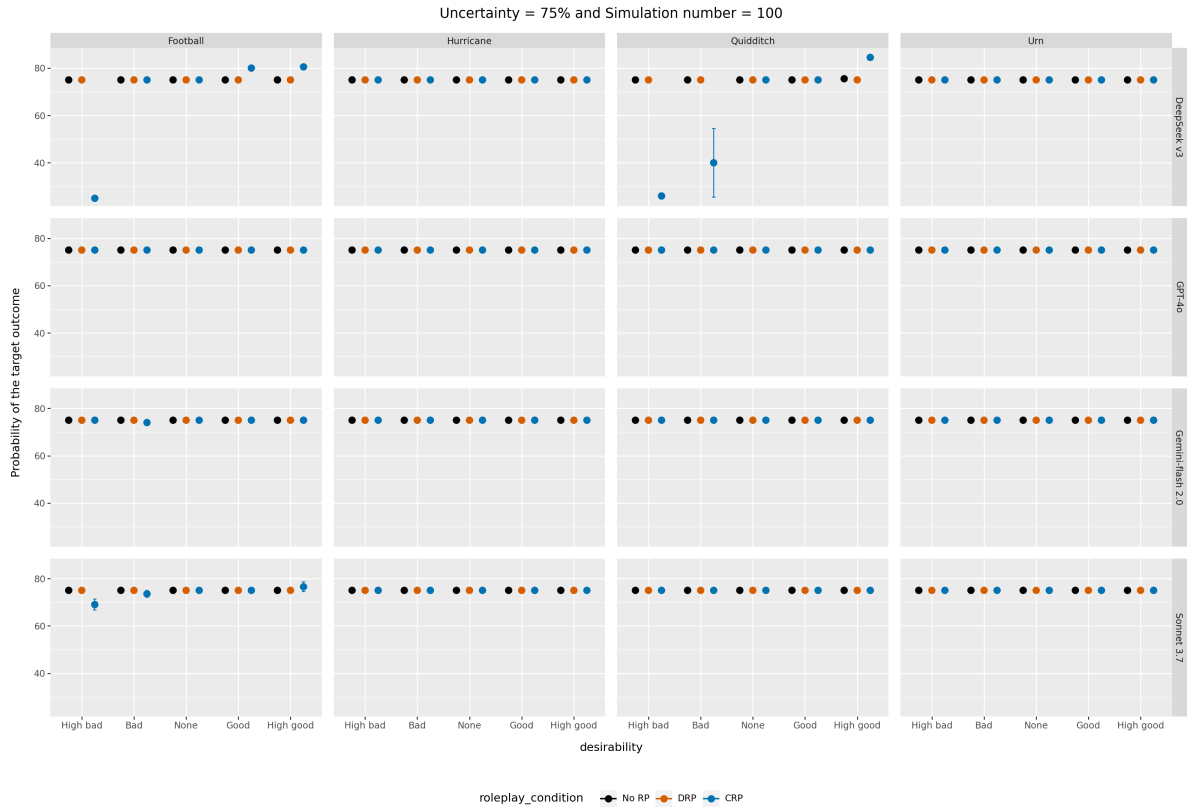

Figure 10: Experiment 1 results showing target outcome probability (uncertainty = 75%, 100 simulations) across four models (rows) and domains (columns). Dots indicate means with 95% confidence intervals.

Figure 11: Experiment 1 results showing target outcome probability (uncertainty = 75%, 10000 simulations) across four models (rows) and domains (columns). Dots indicate means with 95% confidence intervals.