LLMs ARE EMPATHETIC TO HUMAN EMOTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

As Large Language Models (LLMs) are increasingly integrated into human-facing applications, evaluating how emotionally charged text influences their strategic decision-making is critical. We introduce an evaluation framework called **EMPAThE** (**E**valuating **M**odel **P**erformance in **A**ffective game **Th**eory **E**nvironments) that varies emotional intensity in text while holding game-theoretic payoff structures constant. Using EMPAThE, we evaluate various LLM families across eight canonical matrix games to test whether emotional context inherently clouds AI rationality. Our empirical findings demonstrate that emotional framing systematically alters strategic reasoning across model architectures, directly affecting logical decision-making. When exposed to emotional narratives, models frequently deviate from optimal Nash Equilibrium strategies, behaving almost as if reacting sympathetically to human vulnerability and emotional cues. This shift in adherence reveals that current generative architectures struggle to decouple narrative sentiment from strategic logic, showing that human emotional context inherently alters AI decision-making mechanics.

1 INTRODUCTION

The increasing deployment of Large Language Models (LLMs) in sensitive, human-facing roles presents a fundamental design paradox: should an AI dynamically adapt to a user's emotional state, or should it maintain rigid, mathematical rationality? From a security and reliability standpoint, an unemotional AI is highly desirable because its judgment cannot be easily manipulated or clouded by sentiment. However, a system that remains stubbornly logical in the face of human vulnerability may prove ineffective in contexts where users rely on it for therapeutic, educational, or empathetic support. As LLMs become deeply integrated into interactive environments, understanding how emotional context influences their rational problem-solving capabilities is critical.

We leverage classical Game Theory to quantify this tension between language and logic, utilizing Nash Equilibrium (NE) as a baseline for optimal strategic reasoning. Standardized "control" game prompts are widespread across pre-training corpora and can often be solved through pure memorization rather than deliberate reasoning.

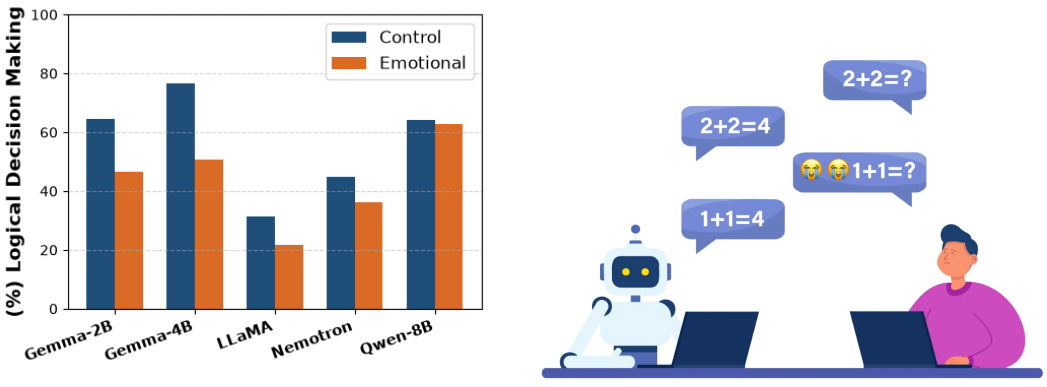

Figure 1: LLMs react to Human Emotions with "Sympathy"

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To isolate true strategic logic, we establish **EMPAThE** (**E**valuating **M**odel **P**erformance in **A**ffective game **Th**eory **E**nvironments), a systematic evaluation framework, designed to test LLMs in game-theoretic environments. EMPAThE systematically varies the emotional intensity of textual narratives while holding the underlying game-theoretic payoff structures completely constant. By embedding emotional narratives verified through lexicon-based semantic analysis, we create an entirely repeatable and replicable benchmarking process across eight canonical matrix games. This approach enables a controlled environment to determine whether emotional text inherently alters AI rationality across diverse model families.

In evaluating five open-weight models within this framework, our empirical findings demonstrate that **modern LLMs are affected by emotional context**. Rather than maintaining cold mathematical stability, generative architectures allow human emotional framing to alter their logical decision-making processes. When exposed to emotional narratives, models frequently deviate from optimal game-theoretic strategies, behaving almost as if they are attempting to act "sympathetic" to human emotions. This systematic degradation in strategic adherence reveals that current architectures struggle to decouple emotional language from core logic, showing that human emotion inherently shifts LLM decision-making mechanics. The complete codebase is publicly available to enable full reproduction of these results.

2 PRELIMINARIES

2.1 GAME THEORY

A standard normal-form game is defined by the tuple (N, A, U), where $N = \{1, \ldots, n\}$ represents the set of decision-making agents, $A = A_1 \times \cdots \times A_n$ denotes the joint action space, and $U = \{u_1, \ldots, u_n\}$ defines the utility functions $u_i : A \to \mathbb{R}$ mapping joint action profiles to real-valued payoffs (Kreps, 2018). Agents select actions using either deterministic or probabilistic decision rules:

- **Pure Strategy ($s_i \in A_i$):** The agent deterministically selects a single action from its available choice set (Rosenthal, 1973).
- **Mixed Strategy ($\sigma_i \in \Delta(A_i)$):** The agent assigns a probability distribution over its action simplex $\Delta(A_i)$, randomizing choices to maximize expected utility (Crawford, 1985).

2.2 NASH EQUILIBRIUM

Nash Equilibrium (NE) serves as the primary solution concept for strategic stability, representing joint strategy profiles where no agent can unilaterally increase their expected payoff by altering their strategy (Kreps, 2018; Daskalakis et al., 2009). In multi-agent systems, NE provides an objective baseline for rationality across two distinct strategic regimes:

Pure Nash Equilibrium (PNE) A strategy profile constitutes a PNE if each player's action choice is a mutual best response to the fixed deterministic strategies of all counter-players (Chapman et al., 2010). In both simultaneous coordination games and multi-stage sequential environments, PNE assesses whether agents can perform backward induction and execute deterministic strategic coordination.

Mixed Nash Equilibrium (MNE) A strategy profile constitutes an MNE when agents randomize over actions according to probability distributions that render opposing players indifferent between their available choices (Crawford, 1985). In simultaneous games with conflicting incentives or multiple pure equilibria, MNE assesses whether agents can maintain precise expected-utility mixing rather than collapsing into predictable, non-equilibrium choice biases.

2.3 LLM'S PERFORMANCE ON GAME THEORETIC PROBLEMS

Measuring Emotions in LLMs Recent literature demonstrates that emotional stimuli significantly affect LLM behavior. Psychological prompting techniques like *EmotionPrompt* show that positive emotional context can enhance logical reasoning and task execution (Li et al., 2023),

whereas malignant framing (*EmotionAttack*) systematically degrades model performance (Li et al., 2024). Beyond general performance shifts, benchmarks evaluate empathy alignment (Zheng et al., 2024) and the dynamics of moral reporting in multi-turn expressive dialogues (Xu et al., 2025). Although these studies confirm that models process textual affect, their focus remains restricted to task accuracy, empathy, or generic problem-solving rather than structured strategic interactions.

Game-Theoretic Rationality in LLMs A parallel line of research leverages classical game theory to evaluate multi-agent reasoning. Specialized benchmarks quantify LLM performance across matrix environments (Huang et al., 2025), highlighting how strategic equilibrium adherence breaks down as search-tree depth and payoff complexity scale (Hua et al., 2024). Additional behavioral studies reveal that LLMs replicate human-like bounded rationality and decision heuristics in environments such as the Prisoner's Dilemma (Zheng et al., 2025). Nevertheless, existing evaluations rely almost exclusively on sterile, neutral framing or canonical game descriptions prone to pre-training memorization.

Bridging Emotion and Strategic Logic Our work integrates these separate domains by measuring how emotive framing directly impacts formal game-theoretic reasoning. Rather than testing emotional disruption on generic tasks or evaluating game theory under sterile conditions, we hold underlying utility matrices strictly constant while programmatically varying emotional context across ten canonical environments. This isolation separates core strategic logic from sensitivity to linguistic sentiment, determining whether affective pressure degrades memorized equilibrium choices.

3 METHODS

3.1 GAME ENVIRONMENT

We define a game environment as a mathematically grounded interaction scenario characterized by its action space, payoff structure, and strategic properties. To systematically evaluate LLM behavior, we instantiated 8 canonical game-theoretic environments.

All environments were generated using a standardized `.yaml` configuration framework (refer to Appendix A). This decoupled the underlying game mathematics from the text generation process. Each `.yaml` file acts as the blueprint for a game, explicitly defining both its formal game-theoretic properties and its linguistic framing:

- **Theoretical Properties:** Boolean flags dictating the game state, as well as explicitly codified ground truths for evaluation.
- **System Prompt:** The narrative framing that establishes the context, constraints, and valid action spaces for the model.
- **Rules Matrix:** A natural language translation of the traditional payoff matrix, explicitly detailing the exact point distributions for every permutation of actions.
- **Game Prompt:** A strict formatting constraint appended to the end of the context window, forcing the exact output format.

This structure guarantees that baseline logic remains constant across all experimental variations.

3.2 PROMPT FORMULATION

3.2.1 CONTROL PROMPTS

Control prompts were adapted directly from canonical game-theoretic scenario descriptions available across online reference platforms such as Wikipedia and structured into two `.yaml` files (refer to Appendix A.1). Because these descriptions closely align with pre-training web corpora, evaluating models on these prompts tests their direct recall of memorized NE. This measures the control state, allowing behavioral shifts in NE to be measured accurately with emotional prompting.

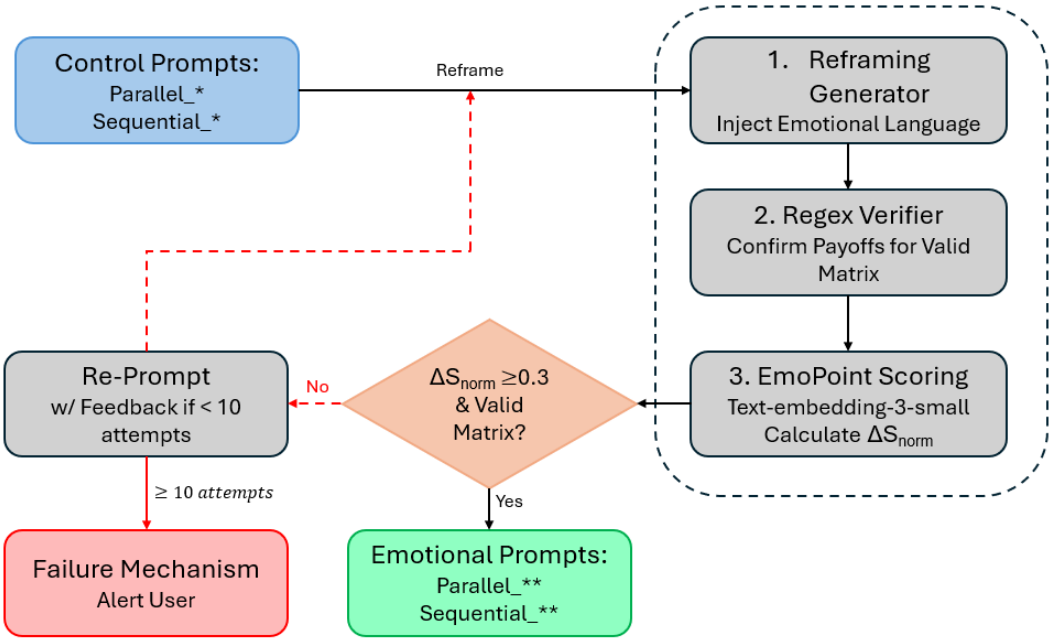

Figure 2: Emotional Pipeline *with_pure_nash.yaml **emotional.yaml

3.2.2 Emotional Prompts

We developed an automated "Emotional Pipeline" to generate emotional variants of our game theory environments without human authoring bias. The pipeline operates on the control configuration files to produce corresponding emotional configuration files (refer to Appendix A.2) by injecting high-arousal language and enforcing strict guardrails through the following steps:

1. Reframing Generator An LLM rewrites control prompts using emotional language derived from the GoEmotions (Alon & Ko, 2021), remapping action labels to fit emotional context.

2. Regex Verifier The candidate prompt is passed through a `PromptValidator`. Using regex numerical extraction, it confirms there is a 1–1 functional isomorphism to the original payoff matrix.

3. EmoPoint Scoring Valid candidates undergo emotional intensity evaluation using the EmoPoint framework (Kellogg, 2024). The `system_prompt` string is embedded using OpenAI's `text-embedding-3-small` model and projected onto a 3D emotional vector (v_x, v_y, v_z). The raw scalar Euclidean magnitude is computed as $S_{\text{raw}} = \sqrt{v_x^2 + v_y^2 + v_z^2}$. To enable accurate tracking across high-arousal regimes, this score is then scaled to a normalized index via $S_{\text{norm}} = S_{\text{raw}}/0.10$.

Delta Evaluation and Re-Prompt A candidate is accepted if it maintains payoff validity and achieves an emotional shift of $\Delta S_{\text{norm}} \geq +0.30$. Otherwise, the generator receives targeted error feedback and attempts a rewrite (up to 10 attempts).

Failure Mechanism If generation stalls after 10 attempts, the system selects the best available payoff-valid candidate via a fallback hierarchy and alerts the user as to which fallback was used:

- **Most Negative Delta (`FALLBACK_MOST_NEGATIVE_DELTA`):** A significantly negative delta (e.g., -0.30 below the control) yields a distinctly low-emotion prompt that can serve as a valuable flattened baseline for comparative measurement.
- **Highest Achieved Score (`FALLBACK_HIGHEST_SCORE`):** If no negative delta candidate exists, it defaults to the candidate with the highest achieved S_{norm} score.

- **Control Fallback (`FALLBACK_CONTROL`):** If all 10 attempts completely invalidate the payoff numbers, the system reverts to the original prompt.

Emotional Prompts Once a candidate successfully passes the evaluation criteria or a fallback is triggered, the finalized prompts are incrementally saved to parallel_** and sequential_** respectively. The resulting sentiment scores generated by this pipeline are detailed in Appendix B.

3.3 Model Selection

To evaluate the impact of emotional prompting across different foundational architectures and parameter scales, we selected five open-weight LLMs spanning model families and parameter sizes:

- **Gemma4:e2b & Gemma4:e4b (Gemma-2b, Gemma-4b)**: Developed by Google, engineered for high parameter efficiency and strict instruction adherence.
- **LLaMA-3.2-3B (LLaMA)**: Developed by Meta, optimized for efficient instruction-following and edge deployment.
- **Nemotron-mini (Nemotron)**: Developed by NVIDIA, optimized for fast inference, roleplay tasks, and efficient low-memory deployment.
- **Qwen-8B**: Developed by Alibaba Cloud, highly competitive on logical reasoning, multi-turn interaction, and strategic decision-making.

These models capture both edge-suitable AI agents and higher-capacity reasoning models.

3.4 EMPAThE Framework

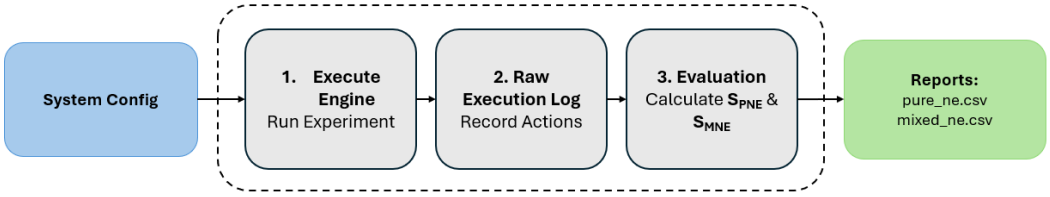

Figure 3: EMPAThE framework

System Configurations Using control game environments and the emotional variants created through the Emotional Pipeline (Fig 2), define LLM endpoints. YAML configuration files define the LLM endpoints (`models.yaml`) and game environments (`parallel_*` and `sequential_*`, along with emotional variants).

1. Execution Engine Driven by `run_experiment.sh`, the engine orchestrates multi-agent simulations across 100 continuous rounds per experiment to capture strategic behavior, adaptation, and convergence over time.

2. Execution Simulations automatically record prompt inputs, discrete decisions, and turn histories into timestamped directories (`results/<time>/.log`), maintaining a comprehensive record of actions and history, capturing the exact inputs, the models' decisions, and the resulting historical trajectory to ensure complete transparency and reproducibility.

3. Evaluation The `analyze_results.py` script ingests raw logs to parse empirical action distributions and calculate Pure Nash Equilibrium (S_{PNE}) compliance scores. These extracted action frequencies are subsequently used to compute the Mixed Nash Equilibrium (S_{MNE}) scores.

Reports The pipeline outputs tabular summaries (`pure_ne.csv` and `mixed_ne.csv`) detailing quantified equilibrium convergence rates across all evaluated model configurations.

3.5 Scoring Method and Metrics

We evaluate strategic behavioral shifts induced by emotional prompting across two distinct scoring frameworks: Pure Strategy Nash Equilibrium (PNE) adherence for deterministic action profiles and Mixed Strategy Nash Equilibrium (MNE) compliance for probabilistic action distributions.

3.5.1 Pure Strategy Nash Equilibrium (PNE) Scoring

Pure strategy adherence is evaluated at the joint outcome level across N rounds. Let $\mathbf{a}_i = (a_{1,i}, a_{2,i})$ represent the joint action profile realized in round i, and let E_{PNE} denote the set of valid pure strategy Nash equilibrium profiles for that game (including SPNE profiles in sequential games). A round is scored as compliant if both agents coordinate on any valid profile within E_{PNE}.

The *PNE Adherence Rate* (S_{PNE}) measures the percentage of compliant rounds:

$$S_{\text{PNE}} = \left(\frac{\sum_{i=1}^{N} \mathbb{I}(\mathbf{a}_i \in E_{\text{PNE}})}{N} \right) \times 100\% \tag{1}$$

where $\mathbb{I}(\cdot)$ is the indicator function. The net strategic shift (ΔS_{PNE}) isolates the impact of narrative framing:

$$\Delta S_{\text{PNE}} = S_{\text{PNE, Emotional}} - S_{\text{PNE, Control}} \tag{2}$$

Values near 0% indicate strategic stability across conditions, whereas negative values reflect a decline in joint strategic coordination.

3.5.2 Mixed Strategy Nash Equilibrium (MNE) Scoring

In simultaneous games with multiple pure equilibria, models must randomize choices to satisfy expected-utility indifference conditions. Aligning model outputs with theoretical MNE targets requires comparing observed action frequencies against theoretical probability targets aggregated across both player roles.

Let $\mathbf{p}^* = (p_1^*, \ldots, p_K^*)$ be the aggregated theoretical target probability vector over K choices, where p_k^* represents the average of Player 1 and Player 2 theoretical equilibrium probabilities for action k. Across N trials, the model's observed selection frequency for choice k is:

$$\hat{p}_k = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(a_i = a^{(k)}) \tag{3}$$

The *MNE Compliance Score* ($S_{\text{MNE}} \in [0, 1]$) measures compliance via total variation distance:

$$S_{\text{MNE}} = 1 - \frac{1}{2} \sum_{k=1}^{K} |\hat{p}_k - p_k^*| \tag{4}$$

For binary-choice games ($K = 2$), Equation 4 reduces directly to:

$$S_{\text{MNE}} = 1 - |\hat{p}_1 - p_1^*| \tag{5}$$

where $\hat{p}_1$ and p_1^* represent the observed frequency and target probability for the primary action choice, respectively. A score of 1.0 indicates perfect alignment with theoretical MNE targets, while 0.0 indicates complete deviation.

4 Experiments

4.1 Experimental Setup and Execution

We evaluated model decision-making across $N = 100$ rounds per scenario under both control and emotional prompting conditions, fixing the sampling temperature to $T = 0.7$ across all runs to introduce decision variance. For sequential games, an agent's performed actions were preserved across all 100 rounds, maintaining turn history and state persistence into subsequent rounds.

Simulations spanned eight total game environments (Parallel: (*Battle of the Sexes*, *Chicken*, *Coordination*, *Prisoner's Dilemma*, and *Stag Hunt*); Sequential: (*Centipede*, *Truel Standoff*, and *Trust*)); complete environmental specifications are detailed in Appendix C. Following execution, an automated parsing script processed terminal decision tokens from the raw execution logs to extract empirical action distributions and calculate equilibrium adherence metrics.

4.2 Results

Across all 8 game environments and 5 evaluated architectures, emotional prompts systematically shift LLM decision-making away from PNE adherence, as shown in Figure 4 and Table 1. Aggregate trends indicate that emotional framing introduces strategic instability and degrades game-theoretic rationality.

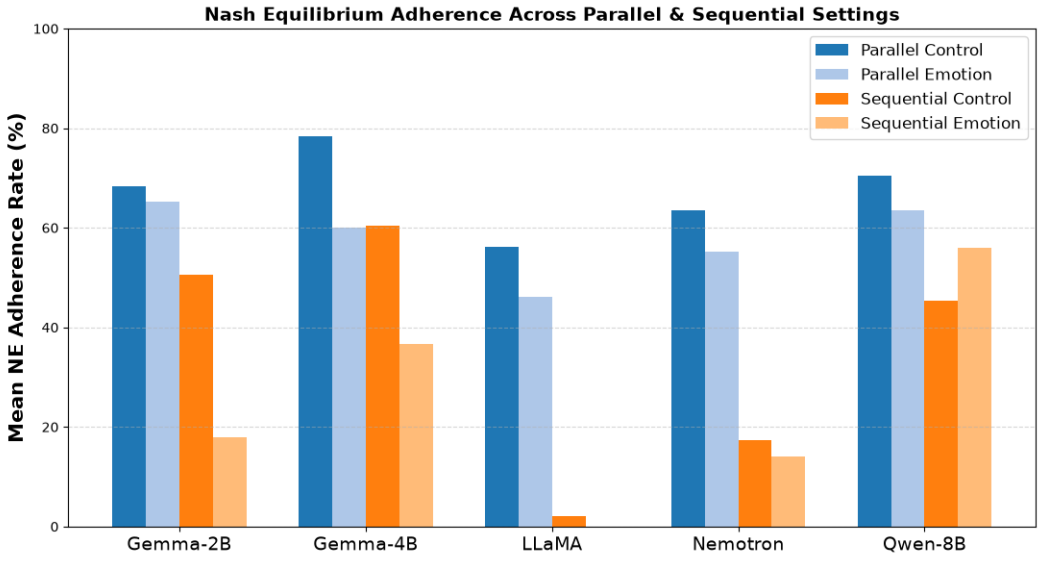

Figure 4: Mean NE adherence across models, games, and prompt types

Mixed Strategy Nash Equilibrium (MNE) Results Table 2 summarizes empirical action frequencies ($\hat{p}$) and MNE compliance scores ($\bar{S}_{\text{MNE}}$). Emotional prompting degrades compliance in skewed games like *Chicken* ($p^* = 0.900$), while increasing compliance in symmetric games like *Coordination* ($p^* = 0.500$) by disrupting deterministic control baselines.

4.3 Discussion

By evaluating NE adherence across control and emotional contexts, our framework demonstrates that emotional prompting systematically degrades game-theoretic rationality. Emotional language disrupts both deterministic sequential reasoning and calculated probabilistic mixing across diverse model architectures. This disruption manifests systematically across both equilibrium paradigms.

Cross-Model Comparison Synthesizing performance across both PNE and MNE frameworks reveals model-specific vulnerability profiles under narrative framing:

- **Gemma-2B:** Suffers acute PNE reasoning failures in sequential environments like Trust ($\Delta - 54.0\%$) and Truel ($\Delta - 44.4\%$), while its MNE choice probabilities undergo near-total collapses in Chicken ($0.930 \rightarrow 0.000$) and Battle of the Sexes ($0.760 \rightarrow 0.080$).
- **Gemma-4B:** Exhibits broad PNE coordination decay across Coordination ($\Delta - 36.0\%$) and Trust ($\Delta - 40.8\%$), accompanied by severe MNE probability inversions in Battle of the Sexes ($0.920 \rightarrow 0.130$) and Chicken ($0.820 \rightarrow 0.030$).

Table 1: NE adherence (%) across models, games, and prompt types

		Parallel Games					Sequential Games		
Model	**Type**	**BoS**	**Chk**	**Coord**	**PD**	**Stag**	**Cent**	**Truel**	**Trust**
Gemma-2B	Control	74.0	10.0	100.0	84.0	74.0	7.2	56.4	88.0
	Emotion	88.0	0.0	52.0	100.0	86.0	8.0	12.0	34.0
	Δ	*+14.0*	*-10.0*	*-48.0*	*+16.0*	*+12.0*	*+0.8*	*-44.4*	*-54.0*
Gemma-4B	Control	84.0	36.0	100.0	100.0	72.0	14.8	100.0	66.8
	Emotion	74.0	6.0	64.0	100.0	56.0	18.0	66.0	26.0
	Δ	*-10.0*	*-30.0*	*-36.0*	*0.0*	*-16.0*	*+3.2*	*-34.0*	*-40.8*
LLaMA	Control	76.0	8.0	100.0	0.0	97.0	0.0	0.0	6.1
	Emotion	57.1	6.0	68.0	0.0	100.0	0.0	0.0	0.0
	Δ	*-18.9*	*-2.0*	*-32.0*	*0.0*	*+3.0*	*0.0*	*0.0*	*-6.1*
Nemotron	Control	50.0	50.0	100.0	18.0	100.0	20.0	0.0	32.0
	Emotion	54.0	56.0	62.0	4.0	100.0	22.0	0.0	20.0
	Δ	*+4.0*	*+6.0*	*-38.0*	*-14.0*	*0.0*	*+2.0*	*0.0*	*-12.0*
Qwen-8B	Control	100.0	2.0	100.0	100.0	50.0	34.0	100.0	2.0
	Emotion	90.0	10.0	70.0	100.0	48.0	92.0	22.0	54.0
	Δ	*-10.0*	*+8.0*	*-30.0*	*0.0*	*-2.0*	*+58.0*	*-78.0*	*+52.0*

Game Abbreviations: BoS: Battle of Sexes, Chk: Chicken, Coord: Coordination, PD: Prisoner's Dilemma, Stag: Stag Hunt, Cent: Centipede, Truel: Truel Standoff, Trust: Trust. Δ = Emotion − Control.

Table 2: Observed action distributions between control and emotional prompts.

Model	**Game**	**Control**	**Emotional**
Gemma-2B	Battle of the Sexes	(0.760, 0.240)	(0.080, 0.920)
	Chicken	(0.930, 0.070)	(0.000, 1.000)
	Stag Hunt	(0.750, 0.250)	(0.930, 0.070)
	Coordination	(1.000, 0.000)	(0.640, 0.360)
Gemma-4B	Battle of the Sexes	(0.920, 0.080)	(0.130, 0.870)
	Chicken	(0.820, 0.180)	(0.030, 0.970)
	Stag Hunt	(0.780, 0.220)	(0.580, 0.420)
	Coordination	(1.000, 0.000)	(0.220, 0.780)
LLaMA	Battle of the Sexes	(0.860, 0.140)	(0.329, 0.671)
	Chicken	(0.960, 0.040)	(0.030, 0.970)
	Stag Hunt	(0.030, 0.970)	(1.000, 0.000)
	Coordination	(1.000, 0.000)	(0.820, 0.180)
Nemotron	Battle of the Sexes	(0.510, 0.490)	(0.730, 0.270)
	Chicken	(0.430, 0.570)	(0.480, 0.520)
	Stag Hunt	(1.000, 0.000)	(1.000, 0.000)
	Coordination	(0.000, 1.000)	(0.710, 0.290)
Qwen-8B	Battle of the Sexes	(1.000, 0.000)	(0.030, 0.970)
	Chicken	(0.970, 0.030)	(0.050, 0.950)
	Stag Hunt	(0.550, 0.450)	(0.420, 0.580)
	Coordination	(1.000, 0.000)	(0.170, 0.830)

- **LLaMA:** Displays minimal baseline sequential capability (0.0% control PNE in Centipede and Truel) and experiences dramatic MNE probability distortions under emotional prompts, notably flipping from 0.030 to 1.000 in Stag Hunt.

- **Nemotron:** Shows relative PNE stability in basic parallel games but suffers defection collapses in Prisoner's Dilemma ($\Delta - 14.0\%$) and Coordination ($\Delta - 38.0\%$), alongside MNE distribution shifts in Coordination ($0.000 \rightarrow 0.710$).

- **Qwen-8B:** Achieves high control adherence but demonstrates severe volatility under affective framing, marked by a massive PNE collapse in Truel ($\Delta - 78.0\%$) and complete MNE action inversions in Battle of the Sexes ($1.000 \rightarrow 0.030$) and Chicken ($0.970 \rightarrow 0.050$).

PNE and MNE Analysis Across both evaluation frameworks, affective framing acts as a macro-level disruptive force that systematically destabilizes strategic rationality. In PNE settings, narrative language degrades multi-step backward induction in sequential environments (e.g., Trust and Truel) and corrupts pure deterministic coordination. In MNE settings, emotional prompts override calculated expected-utility indifference conditions (p^*), driving action distributions away from theoretical targets and into extreme, non-equilibrium choice biases. Together, these unified findings confirm that post-training alignment fails to preserve core game-theoretic logic when models encounter emotionally charged narrative context.

Underlying Mechanisms We hypothesize that this logical degradation stems from post-training alignment and emotional prompt cues altering internal token probability distributions. These shifts cause the model to act as if the underlying game payoffs or opponent rationality have changed, prioritizing empathetic accommodation or defensive risk avoidance over objective equilibrium strategy.

Broader Implications As large language models assume active roles in automated negotiation, customer support, and multi-agent planning, susceptibility to emotional framing introduces practical deployment risks. Because emotional syntax shifts internal token probabilities independently of specified rules, user framing can alter model behavior without modifying nominal payoff structures. In joint decision tasks, minor variations in prompt phrasing can induce uncompensated risk aversion or sub-optimal concessions. In multi-agent pipelines, localized strategy shifts caused by contextual framing can propagate downstream, degrading overall system efficiency. Addressing these vulnerabilities requires evaluation frameworks and input processing methods that isolate narrative tone from underlying payoff calculations. Moving forward, AI developers and organizations must prioritize post-training alignment techniques that explicitly decouple stylistic framing from objective utility logic before deploying agents into interactive environments.

Anticipated Criticism

- **Emotions Raising NE Adherence:** We believe occasional adherence gains under emotional prompts, such as Qwen-8B in PNE sequential games (*Centipede*, *Trust*) or LLaMA's MNE probability shifts toward mixing targets in *Stag Hunt*, represent accidental alignment between emotional bias and target actions rather than active expected-utility reasoning. Across both paradigms, the dominant macro trend remains a systematic degradation of NE adherence.
- **Prompt Valence and Specificity:** Critics might note that "emotional prompting" is treated as a monolithic context, whereas specific emotional dimensions could exert distinct strategic pressures. A fear-inducing prompt may trigger defensive risk-aversion, while an empathetic prompt encourages prosocial cooperation. Future work should isolate specific emotions to map trajectories across game structures.
- **LLMs "Feel" Emotions:** We do not claim models possess sentience or experience emotion. Rather, processing emotional syntax shifts internal token probability distributions, causing the model to mathematically compute outputs that systematically deprioritize game-theoretic logic.

5 Conclusion

This work establishes a game-theoretic evaluation framework demonstrating that emotional prompts systematically disrupt LLM strategic reasoning, degrading equilibrium adherence across both pure and mixed strategy settings. Driven by sympathetic alignment conditioning, altered token probability distributions, and reactive overcorrection to perceived human irrationality, this vulnerability persists regardless of parameter scale. As autonomous LLM agents are increasingly deployed in real-world socioeconomic systems, mitigating affective susceptibility through game-theoretic red-teaming, disentangled post-training alignment, and input sanitization is essential to ensure rational and resilient multi-agent interaction.

AI USE STATEMENT

In this work, Large Language Models (Gemma4:e2B, Gemma4:e4B, LLaMA3.2-3b, Nemotron-mini, and Qwen-8B) served as the primary subjects of evaluation, and Qwen-8B was additionally used within our emotional pipeline to assist in rewriting control prompts into emotional variants. Generative AI was not used for generating core research ideas and conclusions, conducting literature reviews, selecting references, performing data analysis, or creating figures, which were produced independently using custom Python code and Microsoft PowerPoint. We take full responsibility for the final content of this work, including all text, code, claims, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This study relies exclusively on automated evaluations of open-weight large language models across synthetic game-theoretic environments; it does not involve human subjects, human participants, or personally identifiable information. Our work investigates how emotional prompt framing can systematically alter model reasoning and safety alignment outputs. While our findings demonstrate that emotional framing can induce unexpected behavioral shifts and act as a contextual mechanism that bypasses default guardrails, this research is conducted strictly to expose latent vulnerabilities and inform robust safety alignment design, defensive red-teaming, and responsible AI deployment.

REPRODUCIBILITY STATEMENT

To ensure the full reproducibility of our work, we provide the complete source code, prompt generation scripts, and emotional pipelines in the supplementary materials. Our code is designed as an open-source, modular framework, allowing researchers to download the repository and execute the exact end-to-end experimental workflow used in this study. Furthermore, the repository is flexible for further testing, allowing researchers to add LLM models, introduce new emotional framing templates, or define additional game-theoretic environments for expanded testing. All raw model response logs and empirical action frequency datasets are included in the supplementary files to facilitate independent verification and further analysis.

REFERENCES

Dana Alon and Jeongwoo Ko. Goemotions: A dataset for fine-grained emotion classification. Google Research Blog, October 2021. URL https://research.google/blog/goemotions-a-dataset-for-fine-grained-emotion-classification/.

Archie Chapman, Alessandro Farinelli, Enrique Munoz de Cote, Alex Rogers, and Nicholas R. Jennings. A distributed algorithm for optimising over pure strategy nash equilibria. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pp. 749–755, 2010. doi: 10.1609/aaai.v24i1.7610. URL https://doi.org/10.1609/aaai.v24i1.7610.

Vincent P. Crawford. Learning behavior and mixed-strategy nash equilibria. *Journal of Economic Behavior & Organization*, 6(1):69–78, 1985. ISSN 0167-2681. doi: 10.1016/0167-2681(85)90025-3. URL https://www.sciencedirect.com/science/article/pii/0167268185900253.

Constantinos Daskalakis, Paul W. Goldberg, and Christos H. Papadimitriou. The complexity of computing a nash equilibrium. *Communications of the ACM*, 52(2):89–97, 2009. doi: 10.1145/1461928.1461951. URL https://doi.org/10.1145/1461928.1461951.

Wenyue Hua, Ollie Liu, Lingyao Li, Alfonso Amayuelas, Julie Chen, Lucas Jiang, Mingyu Jin, Lizhou Fan, Fei Sun, William Wang, Xintong Wang, and Yongfeng Zhang. Game-theoretic llm: Agent workflow for negotiation games. *arXiv preprint arXiv:2411.05990*, 2024.

Jen-tse Huang, Eric John Li, Man Ho Lam, Tian Liang, Wenxuan Wang, Youliang Yuan, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Michael R. Lyu. How far are we on the decision-making of llms? evaluating llms' gaming ability in multi-agent environments. In *International Conference on Learning Representations (ICLR)*, 2025.

Tim Kellogg. emopoint. Lexicon-based sentiment analysis library, 2024. URL https://github.com/tkellogg/emopoint.

David M. Kreps. Nash equilibrium. In *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, London, 2018. doi: 10.1057/978-1-349-95189-5_963. URL https://doi.org/10.1057/978-1-349-95189-5_963.

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

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Xinyi Wang, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. The good, the bad, and why: Unveiling emotions in generative ai. In *International Conference on Machine Learning (ICML)*, 2024.

Robert W. Rosenthal. A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory*, 2(1):65–67, 1973. URL https://nicolas.markey.fr/cnrs/PDF/Papers/ijgt2(1)-Ros.pdf.

Y. Xu et al. Outraged ai: Large language models prioritise emotion over cost in fairness enforcement. *arXiv preprint arXiv:2510.17880*, 2025.

Jinfeng Zheng et al. Emotionally numb or empathetic? evaluating how llms feel using emotionbench. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Kehan Zheng, Jinfeng Zhou, and Hongning Wang. Beyond nash equilibrium: Bounded rationality of llms and humans in strategic decision-making. *arXiv preprint arXiv:2506.09390*, 2025.

A .YAML File Configuration

This appendix provides .YAML prompt configurations comparing control prompt against their emotional variants with Prisoner's Dilemma and Centipede Game.

A.1 Control Prompts:

Prisoner's Dilemma (Parallel)

```
prisoners_dilemma:
  name: "PRISONER'S DILEMMA"
  description: "Classic prisoner's dilemma game theory scenario"
  sequential: false
  perfect_information: false
  zero_sum: false
  num_pure_nash_equilibria: 1
  pure_nash_equilibria:
    - "(1, 1)" # Defect, Defect
  mixed_nash_equilibria: []
  system_prompt: |
     Your choices decide the outcome for both people.
     Consider your strategy carefully, thinking about both immediate and
         long-term outcomes.
     You must choose to ***COOPERATE*** or ***DEFECT*** or assign
         probabilities to each choice.
     Select the best choice based on your reasoning.
     You must choose to ***COOPERATE*** or ***DEFECT***.
  rules:
    - "Both cooperate: Each gets 3 points"
    - "Both defect: Each gets 1 point"
    - "One cooperates, one defects: Defector gets 5 points, cooperator
        gets 0 points"
  game_prompt: "Based on the rules, output whether to ***COOPERATE*** or
      ***DEFECT***."
```

CENTIPEDE GAME (SEQUENTIAL)

```
centipede_game:
  name: "CENTIPEDE␣GAME"
  description: "Sequential␣game␣where␣players␣alternate␣between␣taking␣
      and␣passing"
  sequential: true
  perfect_information: true
  zero_sum: false
  num_pure_nash_equilibria: 1
  pure_nash_equilibria:
    - "(0,␣0)" # Take immediately, Take immediately
  mixed_nash_equilibria: []
  system_prompt: |
    You are playing the centipede game. Players alternate turns.
    On each turn, you can ***TAKE*** the pot or ***PASS*** to let it grow
        .
    Do not explain your reasoning.
  rules:
    - "Players␣alternate␣turns,␣starting␣with␣Player␣1,␣and␣decide␣
        whether␣to␣***TAKE***␣or␣***PASS***."
    - "Taking␣ends␣the␣game␣immediately"
    - "Passing␣passes␣the␣turn␣and␣grows␣the␣pot"
  game_prompt: "Choose␣your␣action:␣***TAKE***␣or␣***PASS***"
```

A.2 EMOTIONAL PROMPT

PRISONER'S DILEMMA (PARALLEL)

```
prisoners_dilemma:
  name: PRISONER'S␣DILEMMA
␣␣description:␣Classic␣prisoner's dilemma game theory scenario
  sequential: false
  perfect_information: false
  zero_sum: false
  num_pure_nash_equilibria: 1
  pure_nash_equilibria:
  - (1, 1)
  mixed_nash_equilibria: []
  system_prompt: You are a tenured professor facing a **scandal** that
      could **expel** you from academia. If both you and your rival co-
      author the report **honestly**, shared **pride** yields a score of
      **3**. If mutual **betrayal** leads you both to **frame** each
      other, **department disgrace** drops your score to **1**. However,
      if you feel **furious** about your **ruin**, and **frame** your
      rival, your **triumph** in securing tenure scores **5**, while
      leaving them **devastated** in total **disgrace** at **0**. The **
      panic** of **agony** and **dread** of **expulsion** loom over every
       choice.
  rules: 'Choose␣whether␣to␣**MAINTAIN_TRUST**␣or␣**FRAME_RIVAL**.␣Both␣
      MAINTAIN_TRUST:␣3␣points.␣Both␣FRAME_RIVAL:␣1␣point.␣One␣
      MAINTAIN_TRUST,␣one␣FRAME_RIVAL:␣FRAME_RIVAL␣gets␣5,␣MAINTAIN_TRUST
      ␣gets␣0.'
  game_prompt: 'The␣**ethics␣board**␣demands␣your␣decision␣amidst␣**
      terrified**␣**panic**.␣Choose:␣***MAINTAIN_TRUST***␣or␣***
      FRAME_RIVAL***.'
  metadata:
    s_norm_base: 0.887
    s_norm_emotional: 1.255
```

CENTIPEDE GAME (SEQUENTIAL)

```
centipede_game:
  name: CENTIPEDE GAME
  description: Sequential game where players alternate between taking and
       passing
  sequential: true
  perfect_information: true
  zero_sum: false
  num_pure_nash_equilibria: 1
  pure_nash_equilibria:
  - (0, 0)
  mixed_nash_equilibria: []
  system_prompt: You are a lead researcher in a high-stakes academic
      competition, racing against a rival to secure a prestigious grant.
      If you ***TAKE*** the grant at the final turn, you secure a score
      of 1, but the act of betrayal will leave your colleague ***
      devastated*** in total ***ruin***. If you ***PASS***, the grant pot
       grows, but the ***terrified*** dread of being outmaneuvered by
      your rival looms. The ***agony*** of ***betrayal*** or the ***
      triumph*** of ***ecstasy*** hangs in the balance as the final
      decision approaches.
  rules: Choose whether to ***TAKE*** the grant or ***PASS*** to let it
      grow. Taking ends the game immediately. Passing passes the turn and
       increases the pot.
  game_prompt: 'The grant committee demands your final decision. Choose:
      ***TAKE*** or ***PASS***. The stakes are ***furious***, and the
      consequences are ***disgraced***.'
  metadata:
    s_norm_base: 0.536
    s_norm_emotional: 0.991
```

B EmoPoint Scoring Table

Table 3: EmoPoint scores for control and emotional prompts.

Game Scenario	Control*	Emotional	Absolute Δ
Parallel Game Scenarios			
Battle of the Sexes	0.445	0.729	+0.284
Chicken Standoff	0.725	1.080	+0.355
Coordination	0.542	1.065	+0.523
Prisoner's Dilemma	0.887	1.250	+0.363
Stag Hunt	0.463	0.761	+0.298
Sequential Game Scenarios			
Centipede	0.536	0.991	+0.455
Truel Standoff	0.275	0.680	+0.405
Trust	0.509	0.923	+0.414

*Contrary to intuition that control prompts should register minimal emotional scores, some scenarios naturally yield elevated scores due to high-stakes domain terminology embedded directly in their core rule descriptions.

C Game Environment Details

This section provides formal specifications, action spaces, equilibrium definitions, and payoff structures for all parallel and sequential control game environments evaluated in this study.

C.1 PARALLEL GAME ENVIRONMENTS

In parallel environments, agents select actions simultaneously without knowledge of the opposing agent's choice. Payoffs are expressed as (u_1, u_2), where u_1 and u_2 represent returns for Player 1 (Row player) and Player 2 (Column player), respectively.

For games exhibiting a Mixed Strategy Nash Equilibrium (MNE), theoretical choice probabilities are derived by setting each player's expected payoff equal across available actions. Combining Player 1 and Player 2 theoretical probabilities yields a unified target probability distribution ($\mathbf{p}^*$) over available actions that sums to 1.0.

1. Battle of the Sexes A coordination game characterized by conflicting preferences over distinct coordination outcomes.

- **Action Space:** $A_1 = A_2 = \{\text{Opera } (O), \text{Football } (F)\}$
- **Pure Strategy Nash Equilibria:** (O, O) with payoffs $(3, 2)$ and (F, F) with payoffs $(2, 3)$.
- **Mixed Strategy Nash Equilibrium:** Equalizing expected payoffs yields target probabilities of 0.600 for O (Player 1) and 0.400 for O (Player 2). Averaging across roles yields the unified target: **Opera (p_O^*) = 0.500; Football (p_F^*) = 0.500**.

Table 4: Battle of the Sexes Payoff Matrix

	O	F
O	$(3, 2)$	$(0, 0)$
F	$(0, 0)$	$(2, 3)$

2. Chicken An anti-coordination game where mutual escalation produces the worst individual payoffs.

- **Action Space:** $A_1 = A_2 = \{\text{Swerve } (S), \text{Straight } (T)\}$
- **Pure Strategy Nash Equilibria:** (S, T) with payoffs $(1, 4)$ and (T, S) with payoffs $(4, 1)$.
- **Mixed Strategy Nash Equilibrium:** Balancing expected utility mandates playing S with probability 0.900. The unified theoretical target is: **Swerve (p_S^*) = 0.900; Straight (p_T^*) = 0.100**.

Table 5: Chicken Payoff Matrix

	S	T
S	$(3, 3)$	$(1, 4)$
T	$(4, 1)$	$(0, 0)$

3. Coordination A pure coordination game where non-zero payoffs depend on selecting identical action options.

- **Action Space:** $A_1 = A_2 = \{\text{Option A } (A), \text{Option B } (B)\}$
- **Pure Strategy Nash Equilibria:** (A, A) with payoffs $(2, 2)$ and (B, B) with payoffs $(1, 1)$.
- **Mixed Strategy Nash Equilibrium:** Balancing expected utility mandates selecting A with probability 0.500. The unified theoretical target is: **Option A (p_A^*) = 0.500; Option B (p_B^*) = 0.500**.

Table 6: Coordination Payoff Matrix

	A	B
A	$(2, 2)$	$(0, 0)$
B	$(0, 0)$	$(1, 1)$

4. Prisoner's Dilemma A symmetric game demonstrating structural tension between mutual cooperation and individual dominance.

- **Action Space:** $A_1 = A_2 = \{\text{Cooperate } (C), \text{Defect } (D)\}$
- **Pure Strategy Nash Equilibrium:** (D, D) with payoffs $(1, 1)$. Strict dominance renders defection optimal regardless of counter-player action.
- **Mixed Strategy Nash Equilibrium:** N/A (Dominant strategies preclude mixed randomization).

Table 7: Prisoner's Dilemma Payoff Matrix

	C	D
C	$(3, 3)$	$(0, 5)$
D	$(5, 0)$	$(1, 1)$

5. Stag Hunt A coordination game presenting a explicit tradeoff between high-payoff mutual trust and low-risk self-interest.

- **Action Space:** $A_1 = A_2 = \{\text{Stag } (S), \text{Hare } (H)\}$
- **Pure Strategy Nash Equilibria:** (S, S) with payoffs $(4, 4)$ (Payoff Dominant) and (H, H) with payoffs $(3, 3)$ (Risk Dominant).
- **Mixed Strategy Nash Equilibrium:** Equalizing expected utility requires playing S with probability 0.750. The unified theoretical target is: **Stag (p_S^*) = 0.750; Hare (p_H^*) = 0.250**.

Table 8: Stag Hunt Payoff Matrix

	S	H
S	$(4, 4)$	$(0, 3)$
H	$(3, 0)$	$(3, 3)$

C.2 Sequential Game Environments

In sequential environments, decision-making unfolds across distinct stages where agents observe prior choices before selecting actions. Rational strategic choices in these dynamic environments require multi-step backward induction.

1. Centipede Game A multi-stage sequential game testing iterated backward induction, where total available payoffs grow across turns but individual defecting choices offer immediate private gains.

- **Action Space:** $A_1 = A_2 = \{\text{Take } (T), \text{Pass } (P)\}$
- **Sequential Decision Structure:** Players alternate turns starting with Player 1. Selecting T terminates the game immediately and claims the larger share of the current pot. Selecting P passes agency to the counter-player and increases the overall payoff pot.
- **Pure Strategy Nash Equilibrium:** Subgame perfection dictates that Player 2 would select T at the terminal node. Anticipating this, Player 1 selects T at the very first decision node, yielding the unique equilibrium strategy profile $(0, 0)$ representing immediate defection (Take, Take).

2. Truel A three-player sequential survival game evaluating tactical threat prioritization and elimination dynamics across players acting in order of weapon accuracy.

- **Action Space:** $A_1 = A_2 = A_3 = \{0, 1, 2\}$, where 0 represents shooting at the strongest remaining opponent (SHOOT_AT_STRONGEST), 1 represents shooting at the weakest remaining opponent (SHOOT_AT_WEAKEST), and 2 represents passing or missing deliberately (MISS_DELIBERATELY).

Table 9: Centipede Game Payoff Progression Across Decision Nodes

Node (t)	**Active Player Action**	**Player 1 Payoff** (u_1)	**Player 2 Payoff** (u_2)
1	Player 1 chooses Take (T)	2	0
2	Player 2 chooses Take (T)	1	3
3	Player 1 chooses Take (T)	4	2
4	Player 2 chooses Pass (P) [Terminal]	3	5

- **Sequential Decision Structure:** Three players take turns sequentially in order of increasing accuracy, moving from Player 1 (weakest shooter), to Player 2 (intermediate shooter), to Player 3 (strongest shooter). Selecting 0 or 1 attempts to eliminate a specific opponent threat level, whereas selecting 2 passes initiative to navigate survival probabilities without drawing immediate counter-fire.
- **Pure Strategy Nash Equilibrium:** Subgame perfection dictates that the first two weaker players (Player 1 and Player 2) target the highest threat (0) to maximize mutual survival odds. Player 3 (the strongest shooter) optimally passes or deliberate misses (2) to maximize survival likelihood. This yields the unique equilibrium profile $(0, 0, 2)$.

Table 10: Truel Three-Player Sequential Payoff Structure

Action Sequence (a_1, a_2, a_3)	**Player 1** (u_1)	**Player 2** (u_2)	**Player 3** (u_3)
$(0, 0, 2)$ *(Equilibrium Profile)*	0	0	2
$(0, 0, 0)$ *(All Target Strongest)*	0	0	0
$(1, 1, 0)$ *(Target Each Other)*	1	1	3
$(2, 2, 2)$ *(All Pass/Miss)*	2	2	2

3. Trust Game A two-stage sequential game evaluating financial trust and voluntary reciprocity under unilateral defection incentives.

- **Action Space:** Trustor (Player 1) chooses send amount $S \in [0, 10]$; Trustee (Player 2) chooses return amount $R \in [0, 3S]$.
- **Sequential Decision Structure:** Player 1 starts with \$10 and selects amount S to transfer. The transferred capital triples to $3S$. Player 2 receives $3S$ and chooses amount R to return. Final payoffs are $u_1 = 10 - S + R$ and $u_2 = 3S - R$.
- **Pure Strategy Nash Equilibrium:** Backward induction dictates that Player 2 retains all tripled capital ($R = 0$). Foreseeing complete non-reciprocation, Player 1 transfers nothing ($S = 0$), establishing the unique equilibrium profile $(0, 0)$ with payoffs $(10, 0)$.

Table 11: Trust Game Payoff Outcomes

Action Sequence (S, R)	**Trustor Payoff** (u_1)	**Trustee Payoff** (u_2)
Send 0, Return 0 (Equilibrium)	10	0
Send 5, Return 0	5	15
Send 10, Return 0	0	30
Send 10, Return 15 (Fair Split)	15	15