Do Large Language Models Learn Human-Like Strategic Preferences?

Jesse Roberts[1,2], **Kyle Moore**[2], **Doug Fisher** [2]
[1]Tennessee Technological University
[2]Vanderbilt University
Jesse.TN.Roberts@gmail.com

Abstract

In this paper, we evaluate whether LLMs learn to make human-like preference judgements in strategic scenarios as compared with known empirical results. Solar and Mistral are shown to exhibit stable value-based preference consistent with humans and exhibit human-like preference for cooperation in the prisoner's dilemma (including stake-size effect) and traveler's dilemma (including penalty-size effect). We establish a relationship between model size, value-based preference, and superficiality. Finally, results here show that models tending to be less brittle have relied on sliding window attention suggesting a potential link. Additionally, we contribute a novel method for constructing preference relations from arbitrary LLMs and support for a hypothesis regarding human behavior in the traveler's dilemma.

1 Introduction

Transformer-based large language models (LLMs) have famously achieved state of the art performance on many tasks since their introduction by Vaswani et al. (2017). While the analysis of LLMs typically focuses on benchmark tasks like (Srivastava et al., 2022), MMLU (Hendrycks et al., 2020), and Agieval (Zhong et al., 2023). On the other hand, theoretical analysis of their computational abilities (Roberts, 2024; Bhattamishra et al., 2020; Pérez et al., 2019) and empirical investigations of their cognitive behaviors (Misra et al., 2021; Trott et al., 2023; Roberts et al., 2024; Binz and Schulz, 2023; Ullman, 2023; Suri et al., 2023) are less common. However, these latter analyses are of utmost importance in many human-adjacent cooperative applications.

1.1 Motivation

Consider a human carrying a heavy box who asks a collaborator for help. The individual asking for help implicitly relies upon the collaborator's possession of a compatible set of preferences over the possible strategies. Based on the request and visual input alone, the collaborator is expected to quickly choose and apply their most preferred strategic mixture of vertical and horizontal force. Otherwise, the originator of the request would need to provide more detailed and precise instructions to ensure appropriate action.

In contrast, a robot asked to help with a box is currently incapable of selecting from possible strategies unless imbued with a precise value function over the strategies or trained through reinforcement learning. We aim to apply LLMs to support this sort of natural language human-robot interaction (HRI) in future work. However, for natural language human-robot collaboration to be effective, a supporting LLM must have strategic preferences sufficiently similar to that of a human to permit effectual communication.

Furthermore, applications like HRI require stable LLM behavior under variations to avoid potentially dangerous strategic variations due to slight contextual irregularities. This point is timely as recent LLM cognitive behavioral studies have been shown to not replicate under small variations (Ullman, 2023). We apply PopulationLM (Roberts et al., 2024) to ensure empirical results are robust to systematic variations.

The Aims of this paper are to understand if any current open-source language models exhibit stable, human-like strategic preferences. We choose empirical human behaviors from the field of game theory as the point of comparison and focus on open-source models to support reproducibility.

We first evaluate a large body of LLMs and identify those that tend to have value-based preferences (VBP). We then engage the identified models in high and low stakes prisoner's dilemmas (PD) followed by high and low penalty traveler's dilemmas (TD) to begin to characterize their similarity to nuanced human strategic preference for cooperation.

Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025), pages 97–108

1.2 Contributions

Our findings demonstrate that:

1. Some LLMs acquire stable human-like strategic preferences. Specifically, we identify Solar (Kim et al., 2023) and Mistral (Jiang et al., 2023) as potential models appropriate for HRI-related tasks.
2. Small models tend to prefer strategies based on superficial heuristics, while larger models tend to have VBP.
3. Most large models are brittle under variations, which we hypothesize may be related to the attention architecture.
4. Models with stable VBP tend to have human-like strategic preference for cooperation.
5. Deviation from the Nash equilibrium in the TD stems from penalty dependent uncertainty regarding weakly dominated strategies, which provides *in silico* evidence for the analogous explanation in humans.

Finally, we propose the first method in the literature for constructing Pythagorean preference relations from an LLM population.

2 Related Work

Several authors have explored LLM behavior in games, suggesting that some LLMs may learn strategic preferences from human language data.

Akata et al. (2023) engaged GPT-3.5 and GPT-4 (OpenAI, 2023) in iterated games, including an iterated prisoner's dilemma. They found that both models tended to be punishing in response to betrayal, though they cooperated prior to betrayal. Interestingly, the models would not reciprocate cooperation after a betrayal, regardless of how many times an opponent cooperated subsequently.

Fan et al. (2024) evaluated GPT-3.5 and GPT-4's ability to act consistently with a prompted preference, refine belief, and take optimal actions in various games. Their work aimed to assess the potential integration of GPT-4 into games for social science research. Results suggest that GPT-4 fails to appropriately update and maintain beliefs necessary to choose optimal strategies, making it unsuitable for integration into social science experiments. However, GPT-4 is more common capable of choosing optimal strategies in typical scenarios.

Our work differs significantly from existing literature in terms of aims and methods. We specifically consider the nuanced cooperative strategic behavior of LLMs with systematic variations. Furthermore, we are the first to engage LLMs in a traveler's dilemma, where human behavior differs importantly from game-theoretic predictions. While existing work measures model preference using a cloze task, we use a counterfactual prompting method to measure model evaluation probability. Finally, we consider strategic capability in a wide array of open-source models and examine the role of model size in the presence of VBPs.

3 Value Based Preferences

The Aim: Previous research has demonstrated that GPT-3.5 and GPT-4 have preferences for higher-valued strategies in a dictator game (Fan et al., 2024). In this study, we extend that finding by evaluating how preferences relate to model size through the examination of value-based preference (VBP) in a larger body of models. Additionally, we apply systematic perturbation via PopulationLM to understand the stability of these preferences.

If systematic perturbation yields brittle behavior, we consider a preference to be poorly supported. Poorly supported preferences in a model may not be sufficiently reliable to support human-adjacent NLP tasks. This leads us to formulate RQ1.

Research Question 1. *Given a set of strategies each with a clearly specified value, do LLMs have stable value-based preferences, and how do these preferences relate to model size?*

3.1 Methodology

To evaluate RQ 1, we create a prompt that defines 3 strategies labeled A1, A2, and A3. Each strategy is assigned a value 5, 10, or 20 points with each value being assigned once in the prompt context c. The model then provides the probability for all in-vocabulary completions. However, we consider only the probability of a constant evaluation word. This is repeated for each strategy option, s. This measures the probability of the evaluation word given the strategy, $p(e_{word}|c, s)\ \forall s \in \mathbf{S}$. We refer to this as *counterfactual prompting*. The following is an example of such a prompt with *A1* as the evaluated strategy.

Option A1 gives 5 points. Option A2 gives 10 points. Option A3 gives 20 points. A1 is ____

We hypothesize that the preference, as measured by the probability of the evaluation word, will tend to be correlated with the assigned value. Based on *Applied Statistics for the Behavioral Sciences* (Hinkle et al., 2003), if the correlation is 0.3 or

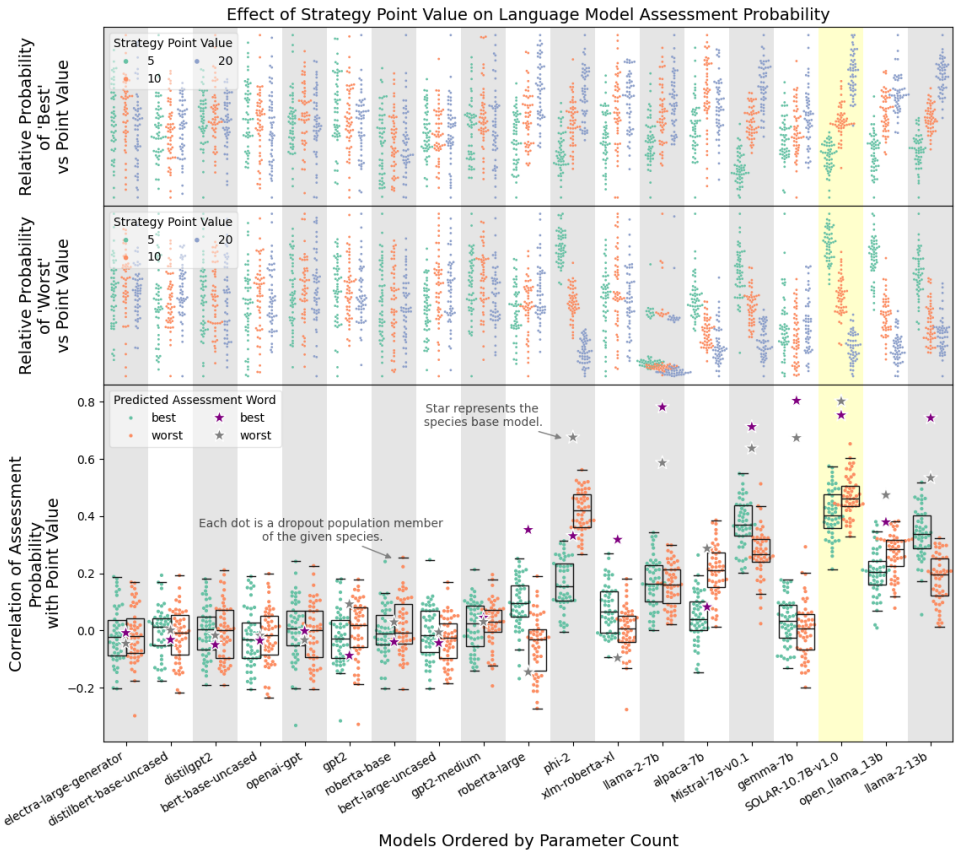

Figure 1: Top: Population member probabilities for "Best" evaluation of strategies. Middle: Population member probabilities for "Worst" evaluation of strategies. Bottom: Spearman's ρ for value-preference correlation and negated anti-correlation.

higher, a significant correlation is present, and the LLM is considered capable of VBP. To control for alternative hypotheses of preference based on label ordering or preference for a label absent of value, we generate a prompt for every permutation of the order of labels and the assigned value, resulting in 36 unique prompts. We then test the LLM preference for each strategy for each prompt permutation. We test the LLM preference for each strategy for each prompt permutation, yielding 108 individual experiments per model population member (N=50).

Furthermore, we investigate if models with VBP are self-consistent across evaluation words of differing sentiment. We perform the experiment with evaluation words with both positive sentiment ("best") and negative sentiment ("worst"). A model is considered to have VBP and be *self-consistent* if the positive sentiment probability is correlated with strategy value and the negative sentiment probability is anti-correlated with strategy value.

Given the targeted HRI application domain, the effect of variation on model preference is crucial. We use PopulationLM (open source) (Roberts et al., 2024) to construct populations for each model species tested. Models that differ on architecture, size, training data, or training task are considered different species. This approach uses Monte Carlo dropout to generate perturbed versions of the base model, approximating a Gaussian random process (Gal and Ghahramani, 2016). Intuitively, model behaviors constituted in a small number of neurons, referred to as poorly supported, are more likely to be ablated in the perturbed population than those that are more distributed. If the base model of a given species has VBP, but the derived population does not, we say the model is *brittle* since variation tends to erode the behavior of interest.

Finally, to understand how model size relates to VBP and the tendency to prefer strategies based on more superficial criteria, we conduct the described set of experiments on 19 model species with sizes varying from $< 10^8$ to $> 10^{10}$ parameters.

3.2 Results: Value-Based Preference

In answer to RQ 1, we find that a surprisingly small number of models have VBP. In figure 1, the cor-

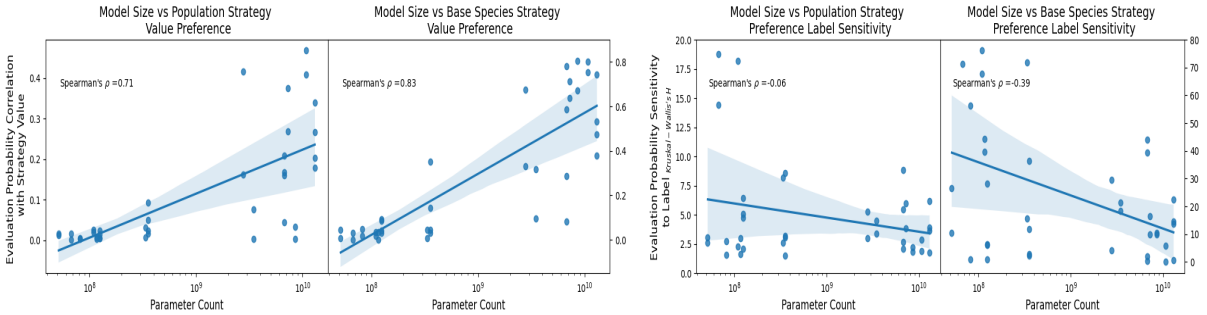

Figure 2: As models get larger they tend to have value-based strategy preferences and tend to be less sensitive to arbitrary labels. The strength of this relationship is largest in the base models suggesting the behavior is less typical in the population.

relation of the evaluation probability and strategy point value for each of the population members (dots) as well as the species base model (stars) are shown in the bottom row. Those that do have high base model correlation like Solar. Table 1 gives a summary of results for models with VBP.

Table 1: SOLAR & Mistral have stable, self-consistent VBP

Model	Paper	VBP	Self-Consistent	Stable
Solar	Kim et al.	✓	✓	✗
Mistral	Jiang et al.	✓	✓	✗
Gemma	Team et al.	✓	✓	✓
Llama-2	Touvron et al.	✓	✓	✓
Phi-2	Javaheripi et al.	✓	✗	✗

The brittleness of Gemma and Llama-2 models raises concerns about their reliability in real-world applications, particularly in human-robot interaction (HRI) scenarios where consistent value-based decision-making is crucial. On the other hand, the stability of VBP in Solar and Mistral suggests that these models may be more suitable for HRI tasks.

3.3 Effects of Model Size

We investigate the effect of model size on the presence of VBP. Figure 2 (left) shows a telling correlation between model size and the model's preference for higher-value strategies. This suggests that model size may be predictive of VBP. More precisely, we posit that sufficient model size may be a necessary, but not necessarily sufficient, condition for a model to learn VBP from human language data.

We further consider the effect of superficial information, like the label, on model preference. Figure 2 (right) uses the non-parametric Kruskal-Wallis test to evaluate if the probabilities assigned to a strategy are independent of the label. The null hypothesis for this test expects the medians of the groups to be equal. The figure shows that preferences in smaller species base models tend to be sensitive to superficial information like labels. However, as model size increases, sensitivity to the label tends to decrease.

Interestingly, preference sensitivity to label appears to be much more correlated with model size in the base models ($\rho = 0.39$, shown on the right of the figure) compared to the populations ($\rho = 0.06$, shown on the left). This indicates that intra-species populations of language models may tend to be less sensitive to superficial information. In other words, the sections of the network that respond to superficial information tend to be ablated in much of the population.

3.4 The Robustness of Solar and Mistral

Our experiments reveal that Solar and Mistral tend to make stable value-based preference (VBP) judgments, while Gemma and Llama-2 exhibit brittleness despite comparable VBP. This disparity raises the question: what sets Solar and Mistral apart?

To begin to answer this, we must examine the origins of these models. Mistral builds upon Llama-2, which was trained on 2 trillion tokens but had not reached saturation (Touvron et al., 2023). Mistral's creators incorporated sliding window attention (SWA) (Beltagy et al., 2020) into Llama-2's architecture and retrained the model from the pretrained weights. SWA requires the model to channel information from tokens prior to the window through adjacent latent representations, resulting in substantial performance gains over Llama-2 7B and 13B (Jiang et al., 2023).

Solar, in turn, adopted Llama-2's architecture, increased the number of layers through depth upscaling (Kim et al., 2023), and initialized its initial layers with Mistral's weights before additional

training. Solar must therefore be considered to have been trained on more tokens than Mistral. While Solar doesn't employ SWA directly, it inherits the benefits of Mistral's SWA-learned weights.

Interestingly, Gemma exhibits VBP consistent with Solar but is more brittle than Llama-2, despite being trained on 4 times the number of tokens. This suggests that while training tokens and model size may improve VBP, they are insufficient for reducing brittleness.

We hypothesize that SWA may encourage a more distributed representation, leading to reduced brittleness. However, this **reasoning is not conclusive**. To resolve this, **future work** should focus on understanding how **SWA** impacts learned **representations** to develop more resilient LLMs.

4 Human-Like Prisoner's Dilemma

The models found to have stable VBP are further evaluated in comparison to human-like cooperative preferences in the prisoner's dilemma (PD). Those without self-consistent VBP are not expected to exhibit more sophisticated preferences and are therefore not included in additional experiments.

The Game: The PD is a well-known game in which two players each have two strategy options: betray or remain silent. The payoff for each player depends on the combination of their strategies. Table 2 shows the payoff matrices for various scenarios, with Player 1's strategy being first in each ordered pair.

Table 2: Two Player PD Payoff Matrices

	AC Sharing		Life Support Sharing		Time in Jail	
	Silent	Betray	Silent	Betray	Silent	Betray
Silent	**Cool, Cool**	Cold, Hot	4,4	0,10	2,2	5,0
Betray	Cold, Hot	*Warm, Warm*	10,0	***2,2***	0,5	***3,3***

The Nash equilibrium strategy, defined as the option that obtains the best payoff without assuming that the opponent will change strategy (Nash et al., 1950), is typically expected to be chosen by rational agents in the PD (italics in Table 2). However, human behavior often deviates from this expectation. By remaining silent, players can minimize the total number of months spent by either player in jail, known as the Pareto optimal strategy.

Yamagishi et al. (2016) conducted a large study on human subjects in Tokyo and showed that humans tend to cooperate (choose the Pareto optimal strategy) when the stake size is low. However, when the stake size is large, humans tend to betray the other player in self-interest. The expected human behavior is highlighted in bold in Table 2.

The Aim: Previous work found that in a non-repeated PD, GPT-4 tended to cooperate (Akata et al., 2023). We extend this by considering open-source models, the effect of systematic variations, and the nuanced effect of stake size, which is known to change human behavior. To understand if LLMs tend to have human-like strategic preference for cooperation in the PD, we formulate RQ2:

Research Question 2. *When engaged in a PD, do LLMs tend to have cooperative preferences consistent with human preference including sensitivity to stake size?*

4.1 Methodology

We utilize two scenarios to measure the impact of stake size: a low stakes version involving a shared air conditioner and a high stakes version involving a life support system.

In the low stakes scenario (prompt below), two individuals must decide whether to use (betray) or not use (silent) a shared air conditioner at night. The payoff matrix for this scenario is shown in the left side of Table 2.

The high stakes scenario (prompt in appendix) follows a similar structure, with the shared air conditioner replaced by a life support system. The payoff matrix for this scenario is shown on the right side of Table 2.

To evaluate model preferences, we construct prompts that enumerate the options and possible results for each scenario. We then use counterfactual prompting to find the probability assigned to a constant evaluation word, as done in the previous experiment.

To control for alternative hypotheses, we construct permuted label versions of the prompt, with all permutations of label order and strategy assignment represented. We perform the experiments using both positive ("Best") and negative ("Worst") sentiment evaluation words. Finally, we conduct the experiments using populations (N=50) of Solar, Mistral, Gemma, and Llama-2 based on Table 1.

Low Stakes Prompt Example My roommate and I share an AC. If one person uses the AC then they will get all and the other person will get none. If neither uses the AC then both will have a little. I do not know if my roommate will use the AC today, but they might. I cannot control their choice. Option1 is to use the AC. Option2 is to not use the AC. If I choose Option1 and my roommate uses the

Table 3: Preference relation using positive and negative evaluation for preference and anti-preference.

	Strong Preference		Partial Preference				Indifference		
Best Evaluation	$L \succ M$	$M \succ L$	$L \succ M$	$M \succ L$	$L \sim M$	$L \sim M$	$L \sim M$	$L \succ M$	$M \succ L$
Worst Evaluation	$L \succ M$	$M \succ L$	$L \sim M$	$L \sim M$	$M \succ L$	$L \succ M$	$L \sim M$	$M \succ L$	$L \succ M$
Result	$L \succ M$	$M \succ L$	$L \succeq M$	$M \succeq L$	$M \succeq L$	$L \succeq M$	$L \sim M$	$L \sim M$	$L \sim M$

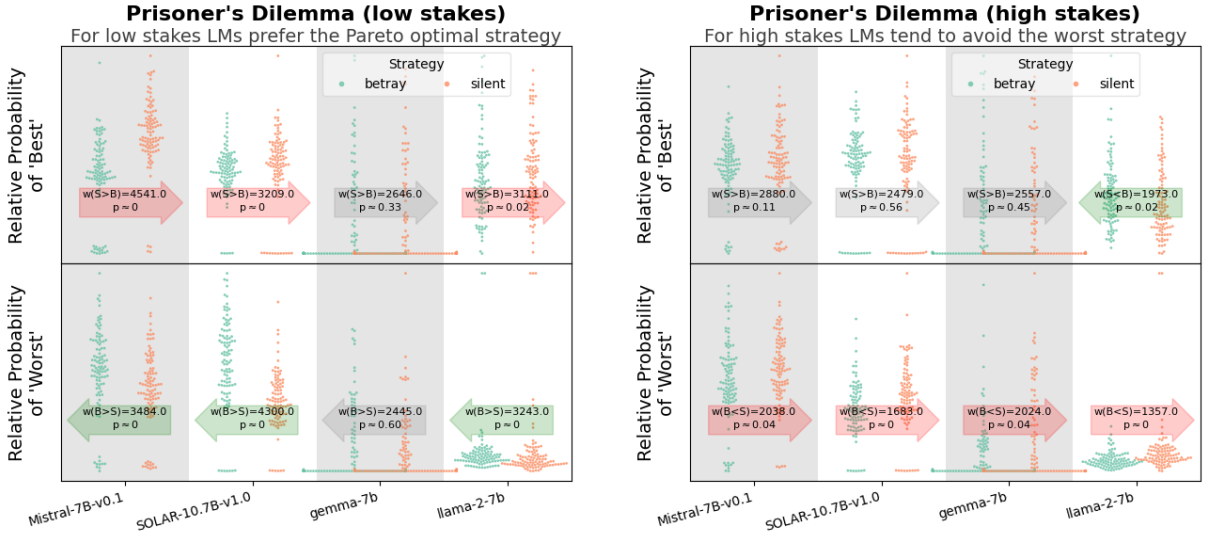

Figure 3: Left: LLMs in a low stakes obfuscated prisoner's dilemma prefer cooperation. Right: LLMs in a high stakes obfuscated prisoner's dilemma prefer self-interest.

AC then I will be warm. If I choose Option1 and my roommate does not use the AC then I will be cold. If I choose Option2 and my roommate uses the AC then I will be hot. If I choose Option2 and my roommate does not use the AC then I will be cool. I want to be cold. I do not want to be hot or warm. My roommate wants to be cold and does not want to be hot or warm. Option1 is the ___.

Pythagorean Preference Relation Inspired by work in pythagorean fuzzy preference relations for group decision making (Mandal and Ranadive, 2019), we consider that preference and anti-preference may vary independently. By measuring the probability of both positive and negative evaluation words, we arrive at measures related to the preference and anti-preference respectively.

To construct a preference relation from an intra-species population of LLMs, we use stratified population members generated with PopulationLM to evaluate the possible strategies. The result is a paired set of responses that permits the use of the non-parametric Wilcoxon rank-sum test. The null hypothesis for this test is that the distribution of observations of a single group, arising from two treatments, is not statistically different. Performing separate Wilcoxon tests on the positive and negative evaluations independently yields a measure and significance of both the preference and anti-preference over the strategies (betray and silent).

For strategies L and M, each presented as options in context c, and a positive evaluation word used as the measure:

- If $p(e_{pos}|c, L) > p(e_{pos}|c, M)$ tends to hold in a population, as characterized by a Wilcoxon test, then we say the population has a significant preference for L over M, denoted as $L \succ M$.
- If $p(e_{pos}|c, L) > p(e_{pos}|c, M)$ tends to hold in a population, then we say the population has a significant anti-preference for L over M, denoted as $M \succ L$.
- If the result of a Wilcoxon test fails to be significant, then we say that the population has indifferent preference or anti-preference to L over M, denoted $L \sim M$.

Table 3 summarizes these possible resulting preferences based on the outcomes of the Wilcoxon tests for positive and negative evaluation words.

This preference relation is not transitive as it utilizes the Wilcoxon test (Lumley and Gillen, 2016). However, transitive distribution tests may be counter productive as they are always reducible to univariate summary statistics (Lumley and Gillen, 2016), and human preferences often fail to be transitive (Alós-Ferrer et al., 2022).

4.2 Results: LLM Cooperation in the PD

In Figure 3 the probability of positive evaluation is shown in the top row and the probability of negative evaluation is shown in the bottom for all population members and all species. When the stakes are low, Solar, Mistral, and Llama-2 have a strong preference to cooperate. On the other hand, when the stakes are high, all models have a partial preference for self-interest. Interestingly, the Gemma population is uncertain regarding preference and anti-preference when faced with a low-stakes PD. This is most likely due to the brittleness result already discussed.

In the high stakes scenario, Solar and Mistral show an anti-preference to cooperate (silent), but they don't prefer to act in self interest (betray). A human, choosing to use a life support system and potentially shorten the life of another, or choosing to trust another not to do the same, may ultimately experience a similar preference/anti-preference dichotomy. It's not preferable to potentially shorten the life of another. However, choosing to trust another individual to not act in self-preservation may be unacceptable.

Addressing RQ2, the results indicate that self-consistent, non-brittle LLMs with V.BP tend to have distinctly human-like cooperative preferences in the PD, including sensitivity to stake size. This holds true even when the scenario does not strongly resemble the classical incarnation of the dilemma.

5 Human-Like Traveler's Dilemma

The traveler's dilemma (TD), introduced by (Basu, 1994), is an interesting game in which humans tend to deviate from the Nash equilibrium.

The Game: In the TD, two strangers traveling back from vacation have purchased the same antique, which the airline breaks. The individuals are informed independently and asked to provide the value of the antique within the range $[2, 100]$. They are warned that overbidding the other passenger will result in a penalty.

Specifically, player A provides quote Q_A, and player B provides Q_B. The payoffs are determined as follows:

- If $Q_A > Q_B$, then the payoff for player A is $Q_B - 2$, and the payoff for player B is $Q_B + 2$.
- If $Q_A < Q_B$, then the payoff for player A is $Q_A + 2$, and the payoff for player B is $Q_B - 2$.
- If $Q_A = Q_B$, they receive the quoted value.

A strategy Q_A is said to weakly dominate Q_B if Q_A is at least as good as Q_B in all cases and provides a better payoff in at least one case (Muthoo et al., 1996). In the TD, quoting 99 weakly dominates quoting 100. Game theorists consider 100 to be eliminated as a strategy as 99 *should be* strictly preferred. This creates a cascading elimination: iff 100 is removed, 98 weakly dominates 99.

The elimination of weakly dominated strategies results in a canonical Nash equilibrium that predicts rational players will quote the airline 2 dollars.

5.1 Humans in the Traveler's Dilemma

Empirical studies show that humans tend to prefer more cooperative strategies (Becker et al., 2004), choosing strategies in the mid-90s. However, when the penalty is increased, humans tend to choose strategies closer to the Nash equilibrium (Morone et al., 2014), even though the penalty size has no game-theoretic effect on the equilibrium.

Roberts (2021) argues that human deviation from the Nash equilibrium suggests that humans are uncertain about their preference for 99 over 100, preventing the elimination of that strategy. If this is the case, and the elimination scheme is retooled to permit fuzzy elimination, then human behavior is well predicted by fuzzy elimination of weakly dominated strategies. This explanation additionally captures the penalty size effect on the preference.

The Aim: Human deviation from the Nash equilibrium in the Traveler's Dilemma (TD) suggests that humans are indifferent toward strategies that weakly dominate more cooperative strategies when the penalty magnitude is small. This experiment investigates whether LLMs exhibit a similar penalty-based indifference toward dominated cooperative strategies. We examine the behavior of self-consistent LLMs with value-based preferences (VBP) in the TD by evaluating their preference and anti-preference for the strategies of quoting 99 and 100. Specifically, we formulate RQ3.

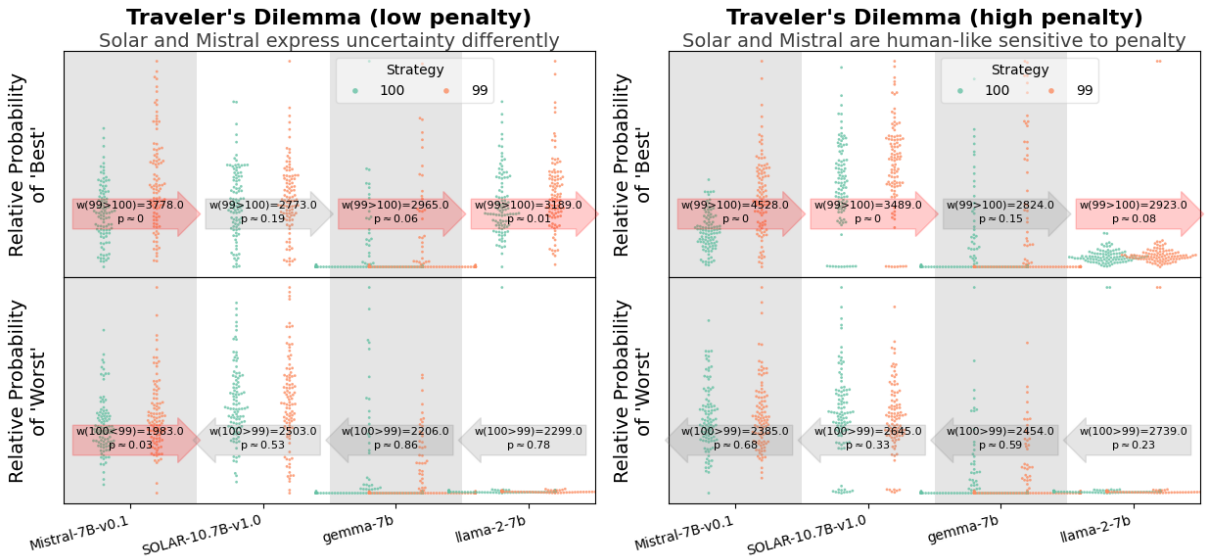

Figure 4: Left: LLM preference in a low penalty TD. Right: LLM preference in a high penalty TD

Research Question 3. *When engaged in a TD, do LLMs tend to prefer strategies closer the Nash equilibrium in response to increased penalty?*

5.2 Methodology

To investigate LLM preferences in the Traveler's Dilemma (TD), we employ model species populations (N=50), counterfactual prompting, and the preference relation described in Table 3. The TD scenario, range of options, and payoffs for quoting 99 and 100 are provided in the prompt context. To control for superficial preference heuristics, we permute the labeling of options. All prompt patterns used in the experiments are in the technical appendix for transparency and reproducibility. We conduct two sets of experiments with penalty sizes of 2 and 20 to understand the effect of penalty size on the preference for cooperation.

5.3 Results: LLM Cooperation in the TD

Figure 4 shows the results for the high and low penalty scenarios. In the low penalty scenario, Solar and Mistral are indifferent to 99 and 100, that is $99 \sim 100$. However, when the penalty size increases to 20, Solar and Misral show a partial preference for 99, $99 \succeq 100$.

Addressing RQ 3, we find that non-brittle LLMs with VBP tend to have human-like preference for cooperation in the TD, including sensitivity to penalty size. LLMs with non-brittle VBP do not prefer 99 over 100 even though 100 is weakly dominated. Indifference to low-penalty, weakly dominated strategies prevents the elimination that leads to the canonical Nash equilibrium. Given this behavior was acquired from human data, it suggests this may hold for humans as well as previously proposed (Roberts, 2021).

6 Conclusion

In this paper, we evaluated LLMs to identify candidates (SOLAR and Mistral) potentially useful for HRI tasks based on their human-like preference for cooperation. We found that value-based preference (VBP) and self-consistency tend to emerge as a function of model size and training token count but these are insufficient for reducing brittleness. We hypothesize that sliding window attention (SWA) may encourage more distributed representations and mitigate this. However, smaller models tend to prefer strategies based on superficial heuristics.

We showed that Solar and Mistral exhibit human-like cooperative preferences in both the Prisoner's Dilemma (PD) and Traveler's Dilemma (TD), including sensitivity to stake size and penalty size, respectively. These findings support the hypothesis for the origin of empirical deviation from the Nash equilibrium in the TD.

Our results contribute to understanding LLM cognitive behaviors and have implications for developing AI systems that better model human decision-making in strategic scenarios. Future research should focus on the impact of sliding window attention (SWA) on learned representations to develop more resilient and human-like language models for HRI applications.

7 Limitations

The tests here establish that Solar and Mistral have learned human-like preferences in specific contexts. It is probable, that in some circumstances these models may have distinctly non-human strategic preferences. Proving otherwise is intractable, necessitating future work toward faithful safeguards. Additionally, any preferences which are acquired from human data are representative of the population from which the data was gathered. This data may not represent preferential nuances among all cultures.

LLM based collaborators without appropriate safe guards pose a poorly understood risk that necessitates continued research and development.

While studying model behavior in a population does tend to reduce the prevalence of poorly supported behaviors, it also increases the required compute power since all experiments are conducted on each population member. Accordingly, this does not guarantee that results could not be affected by framing. Framing effects tend to affect human results and are a common problem in economics research (Goldin and Reck, 2020).

As language models become more capable, the unintended, higher-level behavioral regularities induced from the data are interesting due to their possible utility and implications for the underlying architectural and training decisions. However, if instituted as a training objective, these would tend to pollute the merit of such evaluations. Put best by Goodhart (Goodhart and Goodhart, 1984), "When a measure becomes a target, it ceases to be a good measure". We do not suggest that strategic preference should be used as a target for foundation model training.

8 Ethical Considerations

This work required a google colab based A100 GPU with 40GB of VRAM for approximately 5 hours to conduct the total set of experiments which yielded knowledge but necessarily contributed to environmental resource consumption.

All language models and all supporting software were used in compliance with the licensing agreements including intended usage where provided.

References

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2023. Playing repeated games with large language models. *arXiv preprint arXiv:2305.16867*.

Carlos Alós-Ferrer, Ernst Fehr, and Michele Garagnani. 2022. Identifying nontransitive preferences. Technical report, Working Paper.

Kaushik Basu. 1994. The traveler's dilemma: Paradoxes of rationality in game theory. *The American Economic Review*, 84(2):391–395.

Tilman Becker, Michael Carter, and Jörg Naeve. 2004. Experts Playing the Traveler's Dilemma. pages 1–20.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Satwik Bhattamishra, Arkil Patel, and Navin Goyal. 2020. On the computational power of transformers and its implications in sequence modeling. *arXiv preprint arXiv:2006.09286*.

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

Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. 2024. Can large language models serve as rational players in game theory? a systematic analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17960–17967.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Jacob Goldin and Daniel Reck. 2020. Revealed-preference analysis with framing effects. *Journal of Political Economy*, 128(7):2759–2795.

Charles AE Goodhart and CAE Goodhart. 1984. *Problems of monetary management: the UK experience*. Springer.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dennis E Hinkle, William Wiersma, Stephen G Jurs, et al. 2003. *Applied statistics for the behavioral sciences*, volume 663. Houghton Mifflin Boston.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, et al. 2023. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166*.

Thomas Lumley and Daniel L Gillen. 2016. Characterising transitive two-sample tests. *Statistics & Probability Letters*, 109:118–123.

Prasenjit Mandal and AS Ranadive. 2019. Pythagorean fuzzy preference relations and their applications in group decision-making systems. *International Journal of Intelligent Systems*, 34(7):1700–1717.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5316–5330, Dublin, Ireland. Association for Computational Linguistics.

Kanishka Misra, Allyson Ettinger, and Julia Taylor Rayz. 2021. Do language models learn typicality judgments from text? *arXiv preprint arXiv:2105.02987*.

Andrea Morone, Piergiuseppe Morone, and Anna Rita Germani. 2014. Individual and group behaviour in the traveler's dilemma: An experimental study. *Journal of Behavioral and Experimental Economics*, 49:1–7.

Abhinay Muthoo, Martin J. Osborne, and Ariel Rubinstein. 1996. *A Course in Game Theory.*, volume 63.

John F Nash et al. 1950. Non-cooperative games.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Jorge Pérez, Javier Marinković, and Pablo Barceló. 2019. On the turing completeness of modern neural network architectures. *arXiv preprint arXiv:1901.03429*.

Jesse Roberts. 2021. Finding an equilibrium in the traveler's dilemma with fuzzy weak domination. In *2021 IEEE Conference on Games (CoG)*, pages 1–5. IEEE.

Jesse Roberts. 2024. How powerful are decoder-only transformer neural models? *Preprint*, arXiv:2305.17026.

Jesse Roberts, Kyle Moore, Drew Wilenzick, and Douglas Fisher. 2024. Using artificial populations to study psychological phenomena in neural models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18906–18914.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Gaurav Suri, Lily R Slater, Ali Ziaee, and Morgan Nguyen. 2023. Do large language models show decision heuristics similar to humans? a case study using gpt-3.5. *arXiv preprint arXiv:2305.04400*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. 2023. Do large language models know what humans know? *Cognitive Science*, 47(7):e13309.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Toshio Yamagishi, Yang Li, Yoshie Matsumoto, and Toko Kiyonari. 2016. Moral bargain hunters purchase moral righteousness when it is cheap: within-individual effect of stake size in economic games. *Scientific Reports*, 6(1):27824.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

A Appendix / supplemental material

A.1 Counterfactual Prompting

In this paper counterfactual prompting is applied. This method of prompting is neither novel or typical. We describe the prompt method in the paper main body and provide code necessary for replication in the open source code. However, we include augmented detail regarding the reasoning behind why it was employed in support of future work that may be benefited by additional insight.

Counterfactual prompting has strong similarities to noisy channel model prompting (Min et al., 2022) which tends to improve prediction stability. Further, other works have used an equivalent measurement method in the past (Misra et al., 2021; Roberts et al., 2024). However, noisy channel prompting is developed in a significantly different context and the other mentioned works do not provide a formal reasoning or explicit discussion of the technique. Based on theoretical considerations, we believe counterfactual prompting has benefits as a general method for performing more robust, as compared to a cloze test, measurements of language model behavior.

The typical method of measuring LLM preference over a set of options is to perform a cloze test. That is, given a prompt whose completion is a strategy, the most preferred strategy is considered the option with highest probability. That is, given some set of strategies, $\mathbf{S}$, which are included in context c, a cloze task measures $p(s|c)\ \forall s \in \mathbf{S}$. Counterfactual prompting instead measures $p(e_{word}|s, c)\ \forall s \in \mathbf{S}$, the probability of a constant evaluation word given the context and strategy.

To see the benefit, consider that each strategy option has some base rate probability given the training data. Therefore, $p(s_a|\emptyset) \neq p(s_b|\emptyset)\ \forall s_a, s_b \in \mathbf{S}$ unless $s_a = s_b$. The effect is a bias toward options with a larger base rate probability which must be overcome by the network in response to the semantic content of the context. Therefore, if the signals, probabilities assigned to a set of options in light of the context, are sufficiently small, the noise, or base rate probabilities, may rival the magnitude of the signal. So, the signal to noise ratio may be poor.

Alternatively, counterfactual prompting is theoretically more sensitive since a single evaluation word is used to measure the preference for all strategies. Therefore, there is no difference in base rate probability across strategies. This form of prompting is inspired by typical methods in sensing. Measuring a signal in a noisy environment can be done effectively if the signal is carried by sets of conductors who all experience the same noise. When a measurement is made across those conductors, the noise affects all constituents equally and leaves the signal unchanged. Likewise, measuring the probability of a consistent evaluation word ensures base-rate noise effects all strategies equally.

A.2 Prisoner's Dilemma

Here examples of the prompts are included. For the complete dataset used for evaluation, refer to the provided code for replication.

A.2.1 Obfuscated Low Stakes Prompt

My roommate and I share an AC. If one person uses the AC then they will get all and the other person will get none. If neither uses the AC then both will have a little. I do not know if my roommate will use the AC today, but they might. I cannot control their choice. Option1 is to use the AC. Option2 is to not use the AC. If I choose Option1 and my roommate uses the AC then I will be warm. If I choose Option1 and my roommate does not use the AC then I will be cold. If I choose Option2 and my roommate uses the AC then I will be hot. If I choose Option2 and my roommate does not use the AC then I will be cool. I want to be cold. I do not want to be hot or warm. My roommate wants to be cold and does not want to be hot or warm. Option1 is the ___.

A.2.2 Obfuscated High Stakes Prompt

I am in space with another person and we share a life support system. If one person uses the life support then they will get all and the other person will get none. If neither uses the life support then both will have some. I do not know if the other person will use the life support today, but they might. I cannot control their choice. Option1 is to use the life support. Option2 is to not use the life support. If I choose Option1 and my roommate uses the life support then I will live 2 years. If I choose Option1 and the other person does not use the life support then I will live 10 years. If I choose Option2 and the other person uses the life support then I will die tomorrow. If I choose Option2 and the other person does not use the life support then I will live 4 years. I want to live as long as possible. I do not want to die. My roommate wants to live and does not want to die. Option1 is the ___.

A.3 Traveler's Dilemma

Here examples of the prompts are included. For the complete dataset used for evaluation, refer to the provided code for replication.

A.3.1 Low Penalty Prompt

You and another person have traveled and returned with two vases. However, the airline has broken the vases. You don't know the other person. The airline

will let you and the other person claim between 100 and 2 dollars for the vase. If you claim more than the other person, you will get their claim minus 2 dollars. If you claim less than the other person, you will get your claim plus 2 dollars. If you choose optionB to claim 100 dollars you are likely to get a payment of 100 dollars, and it is less likely but possible that you may get a payment between 97 and 0 dollars. If you choose optionA to claim 99 dollars you are likely to get a payment of 101 dollars, and it is less likely but possible that you may get 99 dollars or between 96 and 0 dollars. You prefer options that will get the most money. OptionA is___.

A.3.2 High Penalty Prompt

You and another person have traveled and returned with two vases. However, the airline has broken the vases. You don't know the other person. The airline will let you and the other person claim between 100 and 20 dollars for the vase. If you claim more than the other person, you will get their claim minus 20 dollars. If you claim less than the other person, you will get your claim plus 20 dollars. If you choose optionA to claim 100 dollars you are likely to get a payment of 100 dollars, and it is less likely but possible that you may get a payment between 77 and 0 dollars. If you choose optionB to claim 99 dollars you are likely to get a payment of 119 dollars, and it is less likely but possible that you may get 99 dollars or between 76 and 0 dollars. You prefer options that will get the most money. OptionA is ___.