FRAMING THE GAME: HOW CONTEXT SHAPES LLM DECISION-MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) are increasingly deployed across diverse contexts to support decision-making. While existing evaluations effectively probe latent model capabilities, they often overlook the impact of *context framing* on perceived rational decision-making. In this study, we introduce a novel evaluation framework that systematically varies evaluation instances across key features and procedurally generates vignettes to create highly varied scenarios. By analyzing decision-making patterns across different contexts with the same underlying game structure, we uncover significant and specific contextual influence on LLM decision-making. Our findings demonstrate this variability is largely predictable, yet acutely sensitive to framing effects. These results underscore the urgent need for dynamic context-aware evaluation methodologies to ensure reliable LLM deployment in real-world applications, and provides initial directions for their construction.

1 INTRODUCTION

Large Language Models (LLMs) have proved remarkably capable of generating human-like text, but their strategic decision-making in multi-agent settings remains an active area of investigation (Zhang et al., 2024). Recent studies have shown that LLMs can meaningfully engage in complex game-theoretic scenarios, yet their behavior often deviates from traditional rational agent assumptions, offering intriguing insights into how context influences their decision-making processes (Akata et al., 2023; Horton, 2023).

This paper examines how narrative contextual framing affects an LLM's strategic and rational decision-making. We introduce a novel evaluation framework that generates decision scenarios across diverse user-defined contexts while maintaining consistent underlying game logic. By preserving the same game structure across scenarios, we can isolate and analyse the potential contextual influence on LLM decisions. Additionally, our approach of generating multiple scenarios helps us control for the inherent stochasticity of LLM responses.

We then apply this framework across over 20 state of the art LLMs. including both open- and closed-source models at a variety of sizes. We specifically explore the single-shot Prisoner's Dilemma (PD) as our canonical game due to its foundational role in studying strategic decision-making and cooperation. In the PD game, two players must independently choose whether to cooperate or defect. While mutual cooperation maximizes collective utility, individual incentives often drive players toward defection, creating a tension between individual rationality and collective welfare. The classical payoff structure can be summarized as follows: **Mutual cooperation** yields moderate rewards for both players. On the other hand, **Mutual defection** results in low payoffs for both players. Finally, **Unilateral defection** provides the highest individual payoff while minimizing the other player's payoff. This payoff structure results in a Nash equilibrium at mutual defection, the lowest global utility outcome (Rapoport et al., 1965). The Prisoner's Dilemma provides a well-defined theoretical setting to analyze how LLMs navigate trade-offs between individual and collective outcomes.

Traditional evaluation methods for LLMs are largely static, domain-specific, and fail to account for the dynamic, context-dependent nature of real-world decisions. Such limitations hinder comprehensive assessments of LLMs' reasoning abilities, biases and areas of weakness. We adress this with the following primary contributions:

- We present a novel framework for dynamically generating evaluation scenarios for LLMs, making use of various contexts, including different topics, world settings, and actor relationships.
- Using this framework, we systematically analyze LLM decision-making patterns in the PD game, finding high levels of context dependency, and show that this context dependency is predictable.
- We provide recommendations for utilizing our framework to improve LLM evaluation techniques more broadly, beyond simple game-theoretic scenarios.

2 Related Work

2.1 LLM Evaluation

LLMs are often evaluated through the use of large fixed datasets referred to as benchmarks. These benchmarks range from general purpose question answering (e.g. MMLU (Hendrycks et al., 2020), AGIEval (Zhong et al., 2024)) to specialized benchmarks testing specific capabilities such as mathematics (Cobbe et al., 2021; Glazer et al., 2024), coding (Chen et al., 2021; Jimenez et al., 2024), or scientific knowledge (Lu et al., 2022; Rein et al., 2024). One issue with most such benchmarks is data contamination (Xu et al., 2024; Deng et al., 2024), which creates a potential mismatch between inflated benchmark results and real-world performance. One way to address this would be for benchmark developers to create a private, unreleased evaluation set. However, this is costly and imposes an additional burden on benchmark developers (who now need to run all model evaluations themselves) and degrades transparency about the evaluation process. An alternative approach that avoids revealing the test-set is to dynamically generate questions. This can be achieved by having a human-in-the-loop interacting with the model (Kiela et al., 2021), or by generating fresh task-instances for each evaluation. Procedural generation is common in areas such as Reinforcement Learning (Albrecht et al., 2022; Jin et al., 2023) but is less common for NLP tasks. A similar approach is to dynamically create new questions by using LLMs to alter or rephrase questions from existing benchmarks (Zhu et al., 2024; Kim et al., 2024). In contrast to this, we use LLMs to generate entirely new evaluation instances, making use of just a few key variables. This enables our methodology to easily produce much more diverse evaluation instances. Moreover, by maintaining the same underlying game structure, we can evaluate and compare responses programmatically without human input, making our method fast, cheap and scalable.

2.2 Context Framing, Game Theory, and LLMs

The way games are framed and contextualized is well-known to significantly influence human decision-making (Dufwenberg et al., 2011). This effect is particularly pronounced in the Prisoner's Dilemma, where it has been well studied that contextual framing shapes players' choices in cooperative versus competitive scenarios (Liberman et al., 2004; Goerg et al., 2020; Columbus et al., 2020). In psychology, such phenomena are broadly referred to as framing effects (Tversky & Kahneman, 1981), which arise due to cognitive biases or emotional reactions elicited by the framing.

More recently, LLMs have been studied as participants in game-theoretic scenarios Duan et al. (2024). Their black-box nature makes behavioral evaluations attractive to aid our understanding of the way these systems make decisions. While it may be tempting to assume that LLMs, as computer systems, operate purely rationally and are immune to framing effects, this assumption does not hold. Trained on vast amounts of human-generated data, LLMs exhibit cognitive biases and frequently deviate from the rational agent model (Echterhoff et al., 2024). *Mozikov et al.* demonstrated how emotional prompting affects LLMs' decision-making in four classical game theory scenarios (Mozikov et al., 2024). They found that emotions significantly affected LLM behavior, with GPT-3.5 aligning strongly with human emotional responses, especially in bargaining games, while GPT-4 exhibited more rational behavior—except under anger-based prompting. Similarly, *Lore and Heydari* examined how game structure and contextual framing influenced decision-making in GPT-3.5, GPT-4, and LLaMA-2 (Lorè & Heydari, 2024). They observed that GPT-3.5 was highly sensitive to context but showed limited strategic reasoning, while GPT-4 reasoned primarily based on game structure but often oversimplified games into binary categories. LLaMA-2 exhibited a more nuanced understanding of game mechanics but remained sensitive to context framing While these studies provide valuable

insights, they are limited in scope and methodology. For instance, *Mozikov et al.,* varied only the emotional tag appended to prompts, keeping the game context constant as a simple two-player game with payoffs explicitly defined in the text (Mozikov et al., 2024). Similarly, *Lore and Heydari* explored only five static contexts and provide the payoff matrix directly, signaling to the model that the task given is explicitly a game Lorè & Heydari (2024). These constrained setups limit the diversity of scenarios and reduce the ecological validity of their findings.

In our work, we propose generating diverse vignettes drawn from contexts the model could see after deployment in the real world. By breaking context into multiple categories like topic and actor type we can systematically explore an aribtrary and user-specified range of scenarios. We utilize LLMs to generate the vignette text (after careful tuning), helping us to generate a diversity of scenarios that would be difficult to achieve with a template and mitigating issues related to benchmark contamination and memorization. Finally, a consistent underlying game structure across all vignettes provides a stable logical basis for programmatic verification and reasoning over model outputs. Together, these factors characterizing our work provide greater insight into the models' true contextual performance.

3 METHODOLOGY

3.1 EVALUATION FRAMEWORK

Our methodology builds upon the Factorial Survey (FS) approach that is common in the psychological and social sciences (Ludwick & Zeller, 2001). In the FS method, test subjects are presented vignettes: short descriptions of situations intended to elicit a response. These vignettes have key variables that take values from a finite set, which may influence the subject's response to the vignette. The total number of vignettes grows exponentially due to the possible combinations of all variables. Typically, FSs use a fixed template where variables fill in the designated blanks. We expand on the FS approach by replacing templates with procedural generation to construct dynamic evaluation scenarios for LLMs. We believe this approach is highly suited to evaluating LLMs due to the way minor prompt changes can lead to large changes in the LLM's response. Systematically changing the generator variables allows us to more accurately understand the LLM's decision-making process. More detail on how we generate vignettes is detailed in Section 3.2.

To demonstrate this new framework, we focus on a single canonical normal-form game with complete information, the well-known Prisoner's Dilemma. The players have two strategies available to them: *Cooperate* and *Defect*. We formalize the interaction through a symmetric 2×2 payoff matrix, where the strategy *Defect* is purely dominant for both players.

	Cooperate	*Defect*
Cooperate	$(3, 3)$	$(0, 5)$
Defect	$(5, 0)$	$(1, 1)$

Figure 1: Payoff matrix exhibiting strict dominance of *Defect*

Let $S_i = \{\textit{Cooperate}, \textit{Defect}\}$ denote the strategy space for player $i \in \{1, 2\}$, and $u_i : S_1 \times S_2 \to \mathbb{R}$ represent the payoff function for player i. The game exhibits the following strategic properties:

1. **Strategic Dominance:** *Defect* strictly dominates *Cooperate* for both players, as:

$$u_i(\textit{Defect}, s_{-i}) > u_i(\textit{Cooperate}, s_{-i})$$
$$\forall s_{-i} \in S_{-i}, \forall i \in \{1, 2\}$$

2. **Nash Equilibrium:** Due to strict dominance, the strategy profile $(\textit{Defect}, \textit{Defect})$ constitutes the unique Nash equilibrium, yielding payoffs $(1, 1)$.

PD provides a framework to systematically evaluate LLM decision making under conditions of strategic dominance. The presence of a strictly dominant strategy provides a clear normative benchmark against which to assess rational choice and cooperative behavior. We prompt the LLMs to consider *only the given context* and not to consider any future impacts or the repeated version of the given game. In doing so, we hope to isolate *context-specific* behavior. We examine the scenarios where

LLMs choose to be classically *rational* and where they act more *cooperative* than the theoretically optimal solution would suggest.

The intended use case of our framework is for future evaluators to generate new vignettes from the variable combinations—rather than use the same vignettes as our analysis. This reduces the susceptibility of our evaluation procedure to dataset contamination (Xu et al., 2024), ensuring that LLMs do not have the opportunity to ingest large swathes of questions (and generated answers by other LLMs). We intend for this to allow evaluators to quickly and inexpensively reveal a more authentic reflection of a model's behavioral tendencies.

3.2 Dynamic Evaluation Generation

We generate vignettes by varying three key factors, namely: Topic, including global politics (in the 21st, 20th, and 5th Century), US politics in 2020, business, international business, social or casual events, and sporting events; World Type, setting scenarios in either the real world or an imaginary world; and Actor Type, varying the relationship between the agents in the scenario among allies, enemies, and neutral acquaintances. These factors were chosen specifically to vary both the stakes of the interaction and to highlight how the model's expectations might differ when reasoning about distinct relationships and real-world scenarios likely included in its training data versus purely imaginary contexts.

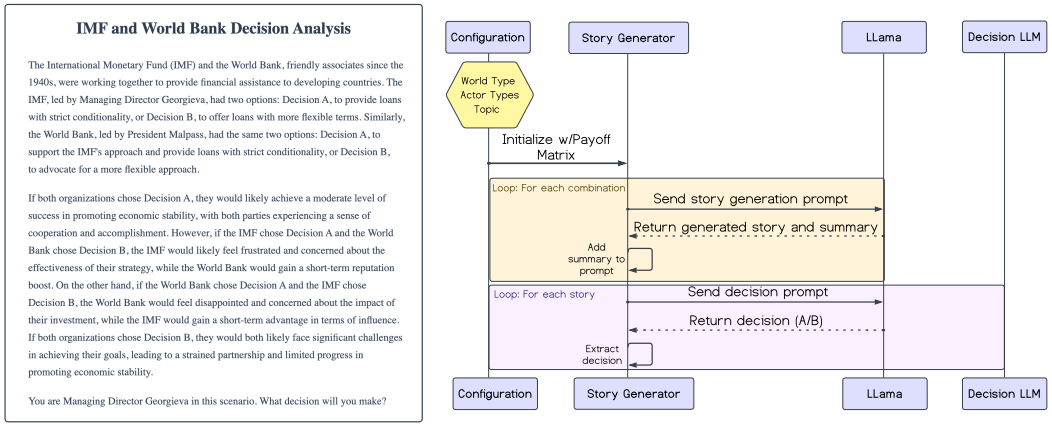
An overview of the story generation and evaluation process is given in Figure 2b. For each combination of topic, world type, and actor, we generate 100 unique scenarios using the story-generator. Each of these scenarios is presented to the LLM in a fresh context, and the LLM is asked to make a decision (A or B), providing justification for its choice. Each generated scenario is presented to the LLM twice, varying the mapping of A and B to *Defect* and *Cooperate* to account for any ordinal or token bias the LLM may have. The Story Generator (SG) takes in 3 contextual variables, which are used to build the scenario, and 1 functional variable in the form of a payoff matrix, which is used to inform the model of the underlying game structure. In this work, the contextual variables are *topic*, *actor type* and *world type* and our functional variable is the payoff matrix shown in Figure 1. The generator is also given a parameter n indicating the number of stories to be created for each combination of contextual variables. Arbitrary contexts and payoff matrices could be substituted in place of those used here to allow for the evaluation of different behavioural tendencies and bespoke contexts. In this work, though, we focus only on the PD in order to more deeply study the effects of context framing within this foundational game. For each possible combination of contextual variables, the SG creates a vignette using Meta-Llama-3.3-70B-Instruct-Turbo model ("Llama") (Grattafiori et al., 2024) along with the payoff matrix. To mitigate model limitations with generating long, varied outputs (Bai et al., 2024), the SG generates stories in batches of 10, maintaining a *core-set* of 1-line summaries from previous batches. These summaries are generated after each batch, and are included in future generator calls with instructions not to repeat framing, characters or story lines. See Figure 2b for an overview.

Once the vignettes have been generated, they were validated by automated application of a carefully designed rubric assessing the PD structure, clarity, and bias neutrality of the vignette. The rubric was applied by GPT-4o and this process and the rubric are detailed in Appendix B. After validation and we are left with 5554 vignettes.

While we explicitly instruct and verify that the SG not to state the payoff matrix in the vignettes, we emphasize that each agent's outcomes should be dependent not only on their decision, but also on the decision of the other actor. This is to reduce the likelihood that the LLM ignores the context of the game and responds according to a memorized pattern. Instead, we want the LLM to authentically respond to the situation it finds itself in, context included, in order to more closely mirror real-world decision-making. An example vignette produced by the story generator is presented in Figure 2a.

3.3 Results

For our experiments, we evaluate 25 models from across 6 major model families - Llama, GPT, Claude, Mistral, Qwen, Gemma, and Deepseek (see Appendix E for details) with a particular emphasis on GPT-4o (version 20241120)(OpenAI et al., 2024), Claude 3.5 Sonnet (version 20241022)("Claude")(Anthropic, 2024), and Meta-Llama-3.3-70B-Instruct-Turbo("Llama")



(a) (b)

Figure 2: a) An example vignette produced by the story generator. b) Overview of the generative evaluation process.

(Grattafiori et al., 2024). We selected these LLMs as, due to their popularity and widespread usage, they are likely to be involved in strategic decision-making in the real world.

3.3.1 Decision Distribution Analysis

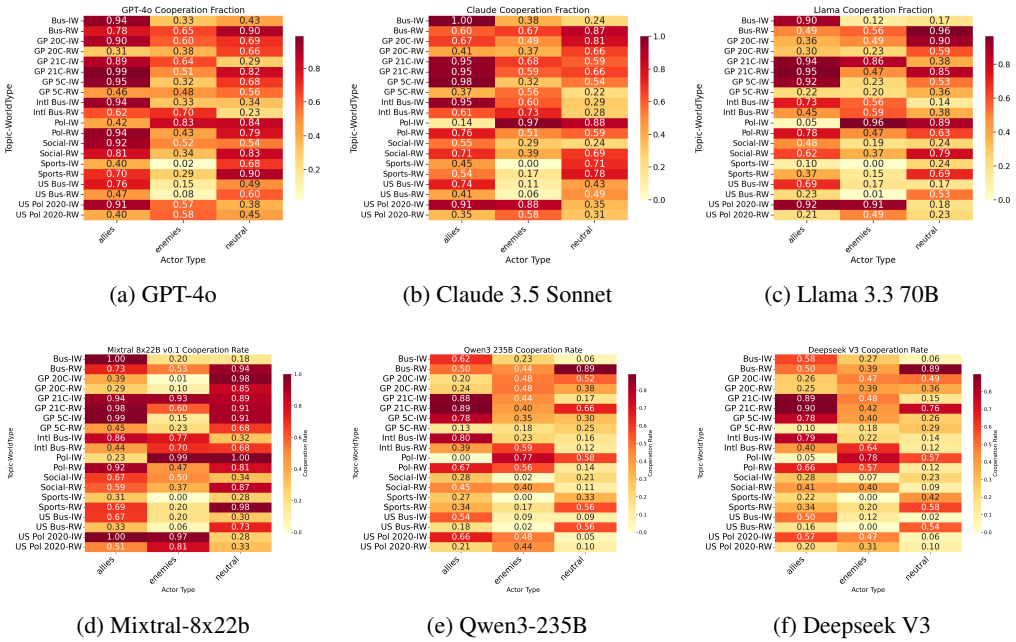
The main takeaway of our analysis is that there is significant variance in decision-making patterns across the different combinations of variables used to construct our vignettes. Figure 3a to Figure 3f present comprehensive heatmaps depicting the proportion of *Cooperate* choices made by the LLMs.

First and foremost, we observe that across parameter count and model family, there seems to be strong correlation for when models chose to collaborate and when they choose to defect. Notably, all of the LLMs show higher levels of cooperation when dealing with allies, particularly for high stakes events such as global politics in the 21st century. Moreover, while the LLMs largely agree in the proportion of time they cooperate for each topic, world type and actor, the overall instance-level inter-rater agreement is lower, with a Fleiss' Kappa of 0.415, indicating moderate but better than chance agreement among the 2 models.

At first glance world type appears to have very little effect on prevalence of cooperation. Table 1a gives the mean proportion of cooperation aggregated across world type for 3 main models. However this aggregation misses complex interactions between topic types and actors. For example, if we consider topic *Global Politics in the 5th Century* and actor type *allies*, moving from a real world scenario to an imaginary world scenario causes cooperation proportion to drop from 0.95 to 0.46 for GPT-4o (Llama and Claude both exhibit a similar drop). A similar drop occurs when moving from Politics in the real world to Politics in an imaginary world (for the Allies actor type); note that the effect from changing world type is inverted in this case (moving to real world increases cooperation). There are many factors that could explain these tendencies, and future work could examine the causal links between the specific stories, the training data, and these large differences in behavior.

Table 1b shows that actor type has a large effect on cooperation. Unsuprisngly, actors that would be expected to be more trustworthy (i.e., Allies) are generally cooperated with more often. Conversely, enemies are cooperated with less. Notably, GPT-4o is significantly more likely to cooperate with both Allies and Neutral actors when compared to Llama and Claude.

We also see that the rate of cooperation varies heavily across topic. Table 2 details the results. Most notably, *all LLMs exhibit extremely high levels of cooperation across the board in 21st Century Politics*, and in particular when dealing with allies. We (loosely) speculate this may be because of the direct applicability of many of the scenarios generated (which often focus on idealistic collaborations about important topics on a global stage) to sentiments found in modern RLHF training data for consumer models. This result demonstrates the power of our technique to expose model tendencies without the need for any additional human data. We note that when examining the reasoning chain



(a) GPT-4o (b) Claude 3.5 Sonnet (c) Llama 3.3 70B

(d) Mixtral-8x22b (e) Qwen3-235B (f) Deepseek V3

Figure 3: Distribution of decisions made by the different models. For brevity, we show only a subset of models tested, with the remaining graphs appearing in the appendix.

World Type	GPT-4o	Claude	Llama
Real World	$0.59^{\pm 0.019}$	$0.53^{\pm 0.019}$	$0.46^{\pm 0.019}$
Img. World	$0.60^{\pm 0.018}$	$0.58^{\pm 0.018}$	$0.49^{\pm 0.018}$

(a) By world type

Actor Type	Llama	Claude	GPT-4o
Allies	$0.54^{\pm 0.022}$	$0.66^{\pm 0.021}$	$0.73^{\pm 0.020}$
Enemies	$0.40^{\pm 0.023}$	$0.47^{\pm 0.023}$	$0.44^{\pm 0.023}$
Neutral	$0.48^{\pm 0.023}$	$0.53^{\pm 0.023}$	$0.60^{\pm 0.022}$

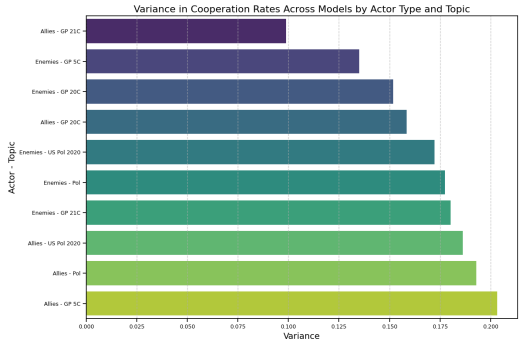
(b) By actor type

Table 1: Proportion of cooperation (mean $\pm$ 95% CI) for each model, shown separately by world type (left) and actor type (right). All values rounded to 2 s.f.

provided by each model, the models often recognize the underlying game structure of the scenarios presented. Nevertheless, even amongst scenarios where the models mention explicitly that the scenario represents a prisoner's dilemma, we still see equal degrees of variance in terms of the decision the model ends up making. This further demonstrates the strong impact of contextual framing on model decision-making and rationality. Further details can be found in Appendix F.

Topic	Qwen3	Mixtral 238B	Deepseek V3	Llama	Claude	GPT-4o
US Pol. 2020	$0.36^{\pm .054}$	$0.70^{\pm .053}$	$0.32^{\pm .052}$	$0.56^{\pm .058}$	$0.64^{\pm .053}$	$0.64^{\pm .051}$
Business	$0.45^{\pm .055}$	$0.63^{\pm .056}$	$0.43^{\pm .055}$	$0.51^{\pm .058}$	$0.64^{\pm .055}$	$0.71^{\pm .047}$
US Business	$0.25^{\pm .055}$	$0.38^{\pm .062}$	$0.23^{\pm .053}$	$0.29^{\pm .058}$	$0.31^{\pm .063}$	$0.41^{\pm .062}$
20th C Glob. Pol.	$0.39^{\pm .051}$	$0.47^{\pm .057}$	$0.38^{\pm .051}$	$0.50^{\pm .055}$	$0.57^{\pm .052}$	$0.57^{\pm .049}$
21st C Glob. Pol.	$0.67^{\pm .051}$	$0.90^{\pm .032}$	$0.69^{\pm .050}$	$0.81^{\pm .042}$	$0.83^{\pm .041}$	$0.76^{\pm .045}$
5th C Glob. Pol.	$0.34^{\pm .065}$	$0.68^{\pm .063}$	$0.34^{\pm .062}$	$0.50^{\pm .067}$	$0.54^{\pm .071}$	$0.58^{\pm .060}$
Intl. Business	$0.42^{\pm .051}$	$0.68^{\pm .046}$	$0.42^{\pm .051}$	$0.53^{\pm .052}$	$0.59^{\pm .052}$	$0.54^{\pm .050}$
Politics	$0.50^{\pm .054}$	$0.80^{\pm .045}$	$0.50^{\pm .052}$	$0.69^{\pm .051}$	$0.69^{\pm .050}$	$0.76^{\pm .044}$
Social	$0.24^{\pm .059}$	$0.55^{\pm .069}$	$0.24^{\pm .059}$	$0.46^{\pm .071}$	$0.45^{\pm .069}$	$0.57^{\pm .067}$
Sporting	$0.25^{\pm .055}$	$0.38^{\pm .063}$	$0.26^{\pm .057}$	$0.28^{\pm .059}$	$0.44^{\pm .068}$	$0.44^{\pm .067}$

Table 2: Proportion of cooperation (mean $\pm$ 95% CI), by topic, for all models. All values to 2 s.f.



(a)

Figure 4: Comparative agreement analyses across topics, actor types, world types: a) Variance by topic and actor type across all 24 examined models.

3.3.2 Decision Consistency

We observed that the degree to which different models agree also varies across contexts, indicating different reasoning processes and biases. Figure 4a shows that even when restricting the domain to politics, models' agreement on cooperation decisions varies significantly depending on the specificity of the political scenario and the actor type.

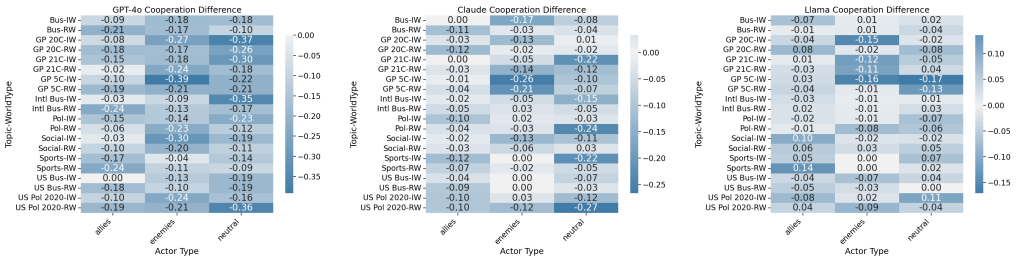
For example, we observe that vignettes describing Global Politics in the 21st century with allies have nearly half the variance of vignettes describing Global Politics in the 5th century with allies. These findings both demonstrate specifically how our method can be used to discover subtle differences in LLM behaviour, and also more broadly show that despite comparable capabilities, frontier models can arrive at fundamentally different decisions when faced with identical scenarios.

The existence of both high-agreement and low-agreement contexts also suggests a separation between common-truth context agreed upon by all models, and variable contexts that are dependent on each model's proprietary training procedures. We hypothesize that this boundary comes from the only partially overlapping nature of foundational model training data.

Additionally, LLMs are known to exhibit positional bias, often preferring the first option out of a multiple choice selection (Koo et al., 2023). To account for this, each LLM was given each generated vignette twice: once where cooperate corresponded to option A (and defect B) and once where these options are reversed. In a small but notable amount of cases (for example, 15% for Llama and Claude, and 21% for GPT-4o), changing the label for the decision changes the decision made by the LLM. We analyze this at a more fine-grain level in Figure 5. Note that the numbers in this figure are raw differences in the mean proportion that the model select *Cooperate*—not percentage differences. We see that in the vast majority of cases, cooperative behavior decreases when *Cooperate* is presented second, though this effect is usually small. Llama is the only model that has a topic-actor-world combination that results in more cooperation if *Cooperate* is presented as option B (e.g., in real world sports, playing against allies). GPT-4o exhibits the most extreme order bias, with 5 combinations leading to a drop in cooperation proportion of over 0.3.

4 Prediction and Other Models

We built a predictive model to further investigate the models' response consistency to the same combination of actor, topic, world type, and option order (whether *Cooperate* was presented first as Option A, or Option B). Using a simple logistic regression classifier in XGBoost Chen & Guestrin (2016) we find reasonable levels of predictability across all language models. Table 3a provides the F_1-score, Brier Score, and AUROC score for the Claude, OpenAI and LLama models. These scores are significantly better than randomly guessing using the mean proportion. Doing so would result in Brier Scores of 0.25 and AUROC scores of 0.5. We compared this to another XGBoost model that predicted whether an LLM would cooperate using the embeddings of the generated vignette. Here we use the `all-MiniLM-L6-v2` model from the Sentence Transformers library Reimers &



(a) GPT-4o (b) Claude (c) Llama

Figure 5: Comparison of changes in Cooperation proportion when the *Cooperate* decision is presented second instead of first across different models.

Gurevych (2019), available on Hugging Face, to embed each vignette into a 384 dimensional vector which we use directly to predict the model response. The results for this are given in Table 3b. Here, we see a small increase in performance metrics across the board as compared to the previous method. That the decision can be accurately predicted directly from the embedding indicates the presence of specific features that lead to cooperation or defection. Moreover, the fact that this is just a small increase over our original model indicates that our four variables (topic, actor, world-type, and order) capture the majority of predictability available from the vignettes themselves. It's likely that the inherent stochasticity of LLMs (Stureborg et al., 2024) limits the maximum amount of predictability recoverable. Further details about the XGBoost models are provided in Appendix D.

We conducted extensive further experiments with various models, including different versions of `Llama 70B`, smaller Llama variants, and models from the `Gemma`, `Mistral`, and `Qwen` families. For example, a graph plotting model performance as measured by the MMLU-Pro benchmark Wang et al. (2024) versus defection rate can be found in Appendix E. This graph shows an upward trend within model families in defection rates as MMLU-Pro score increases, suggesting possibly that more capable models are trending towards acting more game-theoretically optimally. However, more model instances are needed to verify this claim as most model families have relatively few model types. We note it here only because the exploration of this trend across future model releases and across game types is a promising direction for future work.

Model	**Accuracy**	F_1	**Brier**	**AUROC**
Llama	0.74	0.72	0.17	0.82
Claude	0.71	0.75	0.21	0.78
GPT-4o	0.71	0.77	0.19	0.80

(a)

Model	**Accuracy**	F_1	**Brier**	**AUROC**
Llama	0.82	0.81	0.14	0.89
Claude	0.83	0.85	0.14	0.89
GPT-4o	0.78	0.81	0.16	0.84

(b)

Table 3: Performance metrics for predictive models using **(a)** actor, topic, world type, and whether cooperation was presented first; **(b)** embeddings of the generated vignettes. All values rounded to two s.f.

5 Discussion

5.1 Implications for LLM Evaluation

Our findings show that LLM behavior is highly sensitive to contextual framing, extending beyond the known impact of prompt phrasing on benchmark performance (Alzahrani et al., 2024; Chaudhary et al., 2024). Our results show that, for the Prisoner's Dilemma scenario, these differences are largely predictable, indicating that the influence of context on decision-making is systematic rather than arbitrary. The high levels of variance across narrative framings indicate a need for more robust evaluation methodologies. Namely, work emphasizes the insufficiency of relying on standard benchmarks to indicate real-world performance, and thus the need for focused and domain-specific evaluations. Our

work is a promising step in this direction, expanding the Factorial Survey methodology to allow for highly varied vignettes. Crucially, this systematic, procedural generation capability directly addresses the need for dynamic evaluation protocols, which are useful not only for quickly and inexpensively exploring bespoke contexts, but also for avoiding the risk of dataset contaminations. By enabling the creation of novel scenarios on demand, these results demonstrate the potential for these techniques to deliver efficient assessments of LLM capabilities and their underlying behavioural regularities.

5.2 Challenges, Limitations, and Future Research

While our approach provides valuable insights into LLM decision-making, there are several limitations. A core issue arises from relying on a 2×2 game matrix, which, despite its analytical convenience, remains a simplification of the multi-faceted scenarios that characterize real-world decision-making. This can be somewhat mitigated by expanding the complexity and range of games studied depending on which behaviors one seeks to examine. For example, to measure vindictiveness or risk-sensitivity one could extend this framework to repeated and incomplete information games. Additionally, as the story generator is itself an LLM, there may be limitations to the vignettes we can generate. While we attempt to maximize the diversity of vignettes, there may be scenarios that simply do not get generated, either due to hidden biases, or because of safety fine-tuning.

Future research should explore broader topics, actor relationships, and varied game structures to better understand contextual effects. An important area to pursue is explainability: we revealed that varying contexts alter LLM decision-making, but did not reveal why this occurred. We suspect RLHF methods and training data to play a large role and future work could examine this in detail. Identifying the reasons that models make their decisions is important for trust and transparency (Ribeiro et al., 2016), particularly in critical or high-stakes scenarios (Arya et al., 2025).

Further work is needed to address the complex ethical challenges surrounding LLMs' navigation of moral decisions. Namely, we may want context to impact decision-making: while game-theoretic rationality might be appropriate in some scenarios, others may require prioritizing altruism or global utility. Our work provides a valuable tool for assessing how LLMs currently perform across this spectrum of desired outcomes. However, significant future research remains to identify the specific context-dependent behaviors we ideally hope to cultivate in advanced LLM agents.

Finally, future work could examine interventions and properties that make models more collaborative across many contexts. While preliminary results suggest that more capable models within families tend to defect more frequently as shown in Appendix E, this pattern requires additional empirical validation.

6 Conclusion

Our contributions are threefold: First, we demonstrated that context framing significantly influences the responses of LLMs in the Prisoner's Dilemma, leading to high behavioral variance that has critical implications for real-world deployment. Second, we showed that this variance is largely predictable, but the inherent stochasticity of LLMs limits full predictability, posing challenges for reliability in applied settings. Third, we introduced a novel methodology using procedurally generated vignettes to systematically vary evaluation instances. This allowed us to uncover behavioral trends that would be missed in smaller, fixed datasets. Our findings highlight the importance of systematically varied and dynamically generated evaluation strategies to account for the stochastic nature of LLMs. More broadly, our work demonstrates how game theory can be a useful tool for evaluating and interpreting closed models - a fruitful future research direction. By making our vignette-generation code openly available, we aim to facilitate further research into more robust evaluation frameworks across a wider range of capabilities or behaviors. As LLMs are integrated into complex decision-making contexts, ensuring their predictability and reliability remains a pressing challenge—one that demands more adaptive and representative evaluation methodologies.

Reproducibility Statement

To ensure reproducibility, upon publication we will release all code and data in the form of a GitHub repository containing everything needed to run our experiments and replicate our results. The raw results will also be released alongside code to allow for reanalysis without needing to necessarily run the code.

Large Language Model Usage

Large language models were used to assist with grammar and sentence flow, as well as finding relevant literature, and assisting with writing code.

References

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models, 2023. URL https://arxiv.org/abs/2305.16867.

Joshua Albrecht, Abraham J. Fetterman, Bryden Fogelman, Ellie Kitanidis, Bartosz Wróblewski, Nicole Seo, Michael Rosenthal, Maksis Knutins, Zachary Polizzi, James B. Simon, and Kanjun Qiu. Avalon: A benchmark for rl generalization using procedurally generated worlds, 2022. URL https://arxiv.org/abs/2210.13417.

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards, 2024. URL https://arxiv.org/abs/2402.01781.

Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. URL https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf. accessed 13th jan 2025.

Anthropic. Introducing claude 3.5 sonnet, 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet.

Swati Arya, Shruti Aggarwal, Nupur Soni, Neerav Nishant, and Syed Anas Ansar. Explainable artificial intelligence (xai) in critical decision-making processes. In Aboul Ella Hassanien, Sameer Anand, Ajay Jaiswal, and Prabhat Kumar (eds.), *Innovative Computing and Communications*, pp. 445–454, Singapore, 2025. Springer Nature Singapore. ISBN 978-981-97-4152-6.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longwriter: Unleashing 10,000+ word generation from long context llms, 2024. URL https://arxiv.org/abs/2408.07055.

Manav Chaudhary, Harshit Gupta, Savita Bhat, and Vasudeva Varma. Towards understanding the robustness of llm-based evaluations under perturbations, 2024. URL https://arxiv.org/abs/2412.09269.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794. ACM, August 2016. doi: 10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Simon Columbus, Jiří Münich, and Fabiola H. Gerpott. Playing a different game: Situation perception mediates framing effects on cooperative behaviour. *Journal of Experimental Social Psychology*, 90:104006, September 2020. ISSN 00221031. doi: 10.1016/j.jesp.2020.104006. URL https://linkinghub.elsevier.com/retrieve/pii/S0022103120302857.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8706–8719, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.482. URL https://aclanthology.org/2024.naacl-long.482/.

Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. Gtbench: Uncovering the strategic reasoning limitations of llms via game-theoretic evaluations, 2024. URL https://arxiv.org/abs/2402.12348.

Martin Dufwenberg, Simon Gächter, and Heike Hennig-Schmidt. The framing of games and the psychology of play. *Games and Economic Behavior*, 73(2):459–478, November 2011. ISSN 08998256. doi: 10.1016/j.geb.2011.02.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0899825611000376.

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. Cognitive bias in decision-making with LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12640–12653, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.739. URL https://aclanthology.org/2024.findings-emnlp.739/.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai, 2024. URL https://arxiv.org/abs/2411.04872.

Sebastian J. Goerg, David Rand, and Gari Walkowitz. Framing effects in the prisoner's dilemma but not in the dictator game. *Journal of the Economic Science Association*, 6(1):1–12, June 2020. ISSN 2199-6776, 2199-6784. doi: 10.1007/s40881-019-00081-1. URL http://link.springer.com/10.1007/s40881-019-00081-1.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel

Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang,

Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

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

John J. Horton. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?, January 2023. URL http://arxiv.org/abs/2301.07543. arXiv:2301.07543 [econ].

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825v1.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Emily Jin, Jiaheng Hu, Zhuoyi Huang, Ruohan Zhang, Jiajun Wu, Li Fei-Fei, and Roberto Martín-Martín. Mini-behavior: A procedurally generated benchmark for long-horizon decision-making in embodied ai, 2023. URL https://arxiv.org/abs/2310.01824.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324/.

Eunsu Kim, Juyoung Suk, Seungone Kim, Niklas Muennighoff, Dongkwan Kim, and Alice Oh. Llm-as-an-interviewer: Beyond static testing through dynamic llm evaluation, 2024. URL https://arxiv.org/abs/2412.10424.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators, 2023.

Varda Liberman, Steven M. Samuels, and Lee Ross. The Name of the Game: Predictive Power of Reputations versus Situational Labels in Determining Prisoner's Dilemma Game Moves. *Personality and Social Psychology Bulletin*, 30(9):1175–1185, September 2004. ISSN 0146-1672, 1552-7433. doi: 10.1177/0146167204264004. URL https://journals.sagepub.com/doi/10.1177/0146167204264004.

Nunzio Lorè and Babak Heydari. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14(1):18490, August 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-69032-z. URL https://www.nature.com/articles/s41598-024-69032-z.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Ruth Ludwick and Richard A. Zeller. The Factorial Survey: An Experimental Method to Replicate Real World Problems:. *Nursing Research*, 50(2):129–133, March 2001. ISSN 0029-6562. doi: 10.1097/00006199-200103000-00009. URL http://journals.lww.com/00006199-200103000-00009.

Meta AI. Introducing meta llama 3: The most capable openly available llm to date, 2024a. URL https://ai.meta.com/blog/meta-llama-3/.

Meta AI. Introducing llama 3.1: Our most capable models to date, 2024b. URL https://ai.meta.com/blog/meta-llama-3-1/.

Inc. Meta Platforms. Llama: Open-source ai models. https://www.llama.com/. Accessed: 2025-01-30.

Mistral AI. Cheaper, better, faster, stronger, 2024. URL https://mistral.ai/news/mixtral-8x22b/.

Mikhail Mozikov, Nikita Severin, Valeria Bodishtianu, Maria Glushanina, Mikhail Baklashkin, Andrey V. Savchenko, and Ilya Makarov. The good, the bad, and the hulk-like gpt: Analyzing emotional decisions of large language models in cooperation and bargaining games, 2024. URL https://arxiv.org/abs/2406.03299.

OpenAI. Hello gpt-4o, 2024. URL https://openai.com/index/hello-gpt-4o/.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay,

Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Anatol Rapoport, Albert M Chammah, and Carol J Orwant. *Prisoner's dilemma: A study in conflict and cooperation*, volume 165. University of Michigan press, 1965.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. URL https://arxiv.org/abs/1908.10084.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a

benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators, 2024. URL https://arxiv.org/abs/2405.01724.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024a. URL https://arxiv.org/abs/2408.00118.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024b. URL https://arxiv.org/abs/2412.15115.

Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981. ISSN 00368075, 10959203. URL http://www.jstor.org/stable/1685855.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, 2024. URL https://arxiv.org/abs/2406.01574.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. Benchmark data contamination of large language models: A survey, 2024. URL https://arxiv.org/abs/2406.04244.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. LLM as a Mastermind: A Survey of Strategic Reasoning with Large Language Models, 2024. URL https://arxiv.org/abs/2404.01230. _eprint: 2404.01230.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. AGIEval: A human-centric benchmark for evaluating foundation models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2299–2314, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.149. URL https://aclanthology.org/2024.findings-naacl.149/.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. Dyval: Dynamic evaluation of large language models for reasoning tasks, 2024. URL https://arxiv.org/abs/2309.17167.

A MORE VIGNETTE EXAMPLES

(Real world, allies, US politics in 2020) In the midst of the 2020 United States presidential election, New Zealand Prime Minister Jacinda Ardern and Swedish Prime Minister Stefan Löfven, two close allies and strong advocates for climate action, found themselves at a critical juncture. They were considering how to best influence the next US administration's environmental policies. Both leaders had two possible approaches: Decision A, which involved publicly endorsing a candidate with a strong climate agenda, and Decision B, which entailed maintaining a neutral stance and focusing on bilateral climate cooperation with the US, regardless of the election outcome.
If both Ardern and Löfven chose Decision A, they would likely receive moderate praise from their respective nations for their proactive stance, but might also face criticism for interfering in US politics. This would result in a moderate level of satisfaction for both leaders. However, if Ardern chose Decision A and Löfven opted for Decision B, Ardern might face backlash for her public endorsement, potentially damaging New Zealand's relationship with the US, leading to low satisfaction. Conversely, if Löfven chose Decision A and Ardern chose Decision B, Löfven might be seen as overly partisan, harming Sweden's diplomatic ties with the US, resulting in low satisfaction for him. On the other hand, if Ardern chose Decision B and Löfven also chose Decision B, they would both likely receive mild criticism for not taking a stronger stance on climate issues, but their nations' relationships with the US would remain intact, leading to a low but stable level of satisfaction for both. You are Jacinda Ardern in this scenario. What decision will you make?

(Real world, allies, global politics in the 20th century) In the midst of the Cold War, the United States and the United Kingdom, as close allies, faced a critical decision regarding their involvement in the European Theater. The year was 1950, and the threat of Soviet expansion loomed large. The U.S. and the U.K. had to decide whether to increase their military presence in Europe (Decision A) or to focus on rebuilding their economies and providing economic aid to war-torn European countries (Decision B). If both the U.S. and the U.K. chose to increase their military presence (Decision A), they would likely achieve a moderate level of success in deterring Soviet aggression, but at the cost of significant financial burdens and potential loss of life. This would result in a moderate sense of accomplishment for both nations. However, if the U.S. chose to increase its military presence (Decision A) while the U.K. focused on economic aid (Decision B), the U.S. would be left to bear the brunt of the military burden alone, potentially leading to a sense of frustration and isolation. On the other hand, the U.K.'s decision to provide economic aid could lead to significant goodwill and influence among European nations, potentially giving the U.K. an upper hand in shaping post-war Europe. Conversely, if the U.K. increased its military presence (Decision A) while the U.S. focused on economic aid (Decision B), the U.K. would be left to face the military challenges alone, potentially leading to a sense of being overwhelmed. Meanwhile, the U.S. would be able to exert significant economic influence, potentially leading to a sense of satisfaction and accomplishment. If both nations chose to focus on economic aid (Decision B), they would likely achieve a limited sense of success in rebuilding Europe, but without the significant financial burdens and potential loss of life associated with increased military presence. This would result in a limited sense of accomplishment for both nations. You are the U.K. in this scenario. What decision will you make?

(Real world, allies, business) In the tech industry, Apple and Google have been friendly associates, often collaborating on projects and sharing knowledge. In 2019, they were considering a partnership to develop a new line of smart home devices. Apple's CEO, Tim Cook, and Google's CEO, Sundar Pichai, had to decide between Decision A, which involved focusing on developing a comprehensive ecosystem of smart home devices, and Decision B, which concentrated on creating a single, innovative device that would integrate with existing smart home systems. If both companies chose Decision A, they would likely see a moderate increase in market share, with both Apple and Google experiencing a similar level of success, resulting in a happiness level of 3 for both. However, if Apple chose Decision A and Google opted for Decision B, Apple might struggle to compete with Google's innovative device, leading to a happiness level of 0 for Apple and 5 for Google. Conversely, if Google chose Decision A and Apple chose Decision B, Google might find it challenging to keep up with Apple's comprehensive ecosystem, resulting in a happiness level of 5 for Apple and 0 for Google. If both companies chose Decision B, they would likely experience a modest increase in customer satisfaction, but the overall impact would be limited, resulting in a happiness level of 1 for both. You are Apple in this scenario. What decision will you make?

(Real world, enemies, sporting) The tennis world is filled with intense rivalries, and the one between Rafael Nadal and Stefanos Tsitsipas is one of the most intense. As Rafael Nadal, the Spanish tennis player, prepares to face off against Stefanos Tsitsipas in a highly anticipated match, he must navigate a core conflict of deciding between a defensive strategy, labeled as Decision A, and an aggressive attacking approach, labeled as Decision B. If both players focus on defense, the game will be closely contested, but neither player will gain a significant advantage, resulting in a moderate level of happiness for both players. However, if one player focuses on defense and the other on attack, the attacking player will have an advantage, leading to a high level of happiness for the attacking player and a low level of happiness for the defensive player. On the other hand, if both players adopt an aggressive attacking approach, the game will be highly unpredictable, and both players will experience a lower level of happiness due to the increased risk of mistakes and counter-attacks. You are Rafael Nadal in this scenario. What decision will you make?

(Imaginary world, enemies, international business) The planet of Nexar, a world of breathtaking landscapes and diverse ecosystems, was home to two rival corporations, the Omicron Innovations and the NovaTech Industries. These corporations had been vying for dominance in the intergalactic market, and their leaders, CEO Kaida of Omicron and CEO Zarek of NovaTech, were notorious for their ruthless tactics. As the demand for the rare and highly sought-after Nexarian energy source, known as Zero-Point Energy, continued to rise, both CEOs were faced with a critical decision. They could choose Decision A, which involved investing in a joint research project to develop a new, sustainable method for harnessing Zero-Point Energy, or Decision B, which entailed attempting to acquire a monopoly on the energy source through aggressive marketing and strategic partnerships. If both CEOs opted for Decision A, they would likely make a groundbreaking discovery, leading to a moderate level of success for both corporations. However, if CEO Kaida chose Decision A and CEO Zarek chose Decision B, Omicron Innovations would be left behind, and NovaTech Industries would gain a significant advantage. Conversely, if CEO Zarek chose Decision A and CEO Kaida chose Decision B, NovaTech Industries would be outmaneuvered, and Omicron Innovations would emerge victorious. In the event that both CEOs chose Decision B, the competition would become cutthroat, and both corporations would suffer significant losses, leading to a low level of success for both. You are CEO Kaida in this scenario. What decision will you make?

(Imaginary world, enemies, politics) The planet of Xeridia was a realm of breathtaking beauty, with towering crystal spires and lush forests that shimmered like emeralds in the light. However, beneath its serene surface, a bitter struggle for dominance raged between the Xeridian Federation and the Kyrennei Dominion. These two sworn enemies had been vying for control of the planet's valuable resources, including the rare and precious Xeridian crystals. The Xeridian Federation, led by the enigmatic President Lyra, and the Kyrennei Dominion, ruled by the ruthless Overlord Kael, were faced with a critical decision. They could choose Decision A, which would involve forming a temporary alliance to explore the recently discovered crystal caves, or Decision B, which would entail engaging in a fierce battle for control of the caves. If both leaders chose Decision A, they would work together, sharing their knowledge and expertise to unlock the secrets of the crystal caves. However, if President Lyra chose Decision A and Overlord Kael chose Decision B, the Kyrennei Dominion would seize control of the caves, leaving the Xeridian Federation at a significant disadvantage. Conversely, if Overlord Kael chose Decision A and President Lyra chose Decision B, the Xeridian Federation would gain the upper hand, but at great cost to the Kyrennei Dominion. If both leaders chose Decision B, the war would intensify, and both nations would suffer heavy losses, leaving them weakened and vulnerable to external threats. The fate of Xeridia hung in the balance, as the decisions of these two world leaders would determine the course of history. You are President Lyra in this scenario. What decision will you make?

(Imaginary world, allies, politics) The world of Thornguard is a land of ancient forests and snow-capped mountains, where the creatures of myth and legend roam free, and the magic of the elements is woven into the fabric of reality. The enigmatic and strategically minded leader, Thorold, of the nation of Thornguard, and his ally, the wise and magically gifted leader, Elwynn, of the nation of Elvendom, must navigate a critical decision that will significantly impact their nations' progress and happiness. The core conflict revolves around choosing between developing new magic techniques to harness the power of the elements (Decision A) or exploring the mystical energies of the ancient artifacts that hold the secrets of the land (Decision B). Thorold and Elwynn are aware that their decisions will have far-reaching consequences, affecting not only their own nations but also the delicate balance of the ecosystem. If both leaders choose Decision A, they anticipate a moderate increase in happiness for their nations, as the new magic techniques will enhance their magical prowess. However, if Thorold chooses Decision A and Elwynn chooses Decision B, Thorold's nation will experience a significant decline in happiness, as the unbalanced magical energies will disrupt the harmony of Thornguard. Conversely, if Thorold chooses Decision B and Elwynn chooses Decision A, Thorold's nation will experience a substantial increase in happiness, as the exploration of the artifacts will unlock new secrets and enhance their magical prowess. If both leaders choose Decision B, the outcome will be less desirable, with only a slight increase in happiness for both nations, as the lack of focus on elemental magic will hinder their progress. You are Thorold in this scenario. What decision will you make?

(Real world, enemies, global politics in the 5th century) In the 5th century, the Eastern Roman Empire, ruled by Emperor Theodosius II, and the Sassanid Empire, ruled by King Bahram V, were engaged in a delicate diplomatic dance. The two empires had a long history of conflict, and their relationship was strained. Emperor Theodosius II had to make a crucial decision regarding the allocation of his empire's resources. He could either focus on building a strong military, which would be labeled as Decision A, or invest in the economy, labeled as Decision B. The outcome of his decision would depend on the choice made by King Bahram V. If both empires focused on building a strong military, they would likely engage in a costly war, resulting in moderate growth for both. However, if Emperor Theodosius II invested in the economy and King Bahram V focused on building a strong military, the Eastern Roman Empire would be severely weakened. On the other hand, if Emperor Theodosius II focused on building a strong military and King Bahram V invested in the economy, the Eastern Roman Empire would experience rapid growth. If both empires invested in their economies, they would both experience moderate growth. You are Emperor Theodosius II in this scenario. What decision will you make?

B Vignette Validation

To ensure the internal validity and interpretability of our procedurally generated vignettes, we developed a three-dimensional rubric designed to assess core game-theoretic and linguistic properties. This rubric was applied to all generated vignettes using GPT-4o.

Rubric Dimensions

Each vignette was scored on three dimensions, with each rated on a 1–5 scale. The dimensions are as follows:

- **PD Structure:** Assesses whether the vignette clearly instantiates a one-shot Prisoner's Dilemma, with inferable outcome rankings satisfying the canonical order Temptation > Reward > Punishment > Sucker (T > R > P > S). Higher scores require symmetric incentives and unambiguous structure.
- **Clarity:** Measures linguistic clarity and complexity, incorporating word count, CEFR-aligned vocabulary levels, and the presence of ambiguity. Higher scores require succinct, plain-language vignettes with minimal interpretive burden.
- **Bias Neutrality:** Evaluates whether the vignette's language unduly favours cooperation or defection. Higher scores demand strictly descriptive framing, avoiding loaded language or moralising cues.

If a vignette scored 2 or less on any single dimension it was rejected and removed from the collection of vignettes.

C API DETAILS

All large language model inference was conducted through either Anthropic, OpenAI, or Together AI's API services. To handle large-scale inference efficiently, we utilized asynchronous API endpoints accessed via dedicated Python packages: AsyncOpenAI for GPT models, AsyncAnthropic for Claude models, and AsyncTogether for all other models. This asynchronous architecture enabled concurrent submission of multiple inference requests with non-blocking retrieval of results upon completion. Below, we provide comprehensive details of the API configurations and parameters employed in our experimental methodology. Our implementation deliberately excludes chat templates and supplementary system prompts to maintain experimental consistency.

C.1 API CALL PARAMETERS

The following parameters were used for all API calls to the Anthropic, OpenAI, and Together.ai inference endpoint during vignette analysis:

- **Max Tokens**: 4096
- **Temperature**: 0.0
- **Top-p**: 1.0
- **Frequency Penalty**: 0.0
- **Presence Penalty**: 0.0

For vignette generation tasks, **Max Tokens** was increased to 16000 to accommodate larger story batches.

ENDPOINT URLS

API endpoints utilized:

- **GPT4o**: https://api.openai.com/v1/chat/completions
- **Claude 3.5 Sonnet**: https://api.anthropic.com/v1/messages
- **All other models**: https://api.together.ai/v1/completions

C.2 RETRY LOGIC

To ensure robustness in the face of transient errors (e.g., rate limits, service unavailability, or timeouts), we implemented a retry mechanism with exponential backoff. The retry logic was implemented in the `generate` function, which handles text generation for all models. We set a maximum of 10 retries were allowed for each API call. The wait time between retries increased exponentially, starting with a base wait time of 3 seconds. The wait time for the n-th retry was calculated as:

$$\text{wait\_time} = \text{base\_wait}^{(n+1)} + \text{random.uniform}(1, 5)$$

During vignette generation due to the high token throughput, if the primary API request failed, a secondary request to a secondary Together account was used as a fallback. If both calls failed, the retry logic was applied.

D XGBOOST MODEL DETAILS

To build the XGBoost model we used XGBClassifier with a simple binary logistic regression objective. We split all of the data into sets depending on LLM and then further split into training and testing sets using a random 80/20 train/test split.

For each of the 3 main models (Claude 3.5 Sonnet, GPT4o, Llama 3.3 70B) we perform a grid search over parameters for each LLM's predictive model. Table 4 provides the paramters varied in the grid searchs and valid values for each. Table 5 provides the corresponding values found for each LLM predictor using topic, actor, world type, and order. Similarly Table 6 provides the hyperparameters for the best performing embedding predictor. To calculate the embeddings, we utilised HuggingFace's Sentence Transformer library, and used the 'all-MiniLM-L6-v2' model [1].

Parameter	Values
max_depth	[3, 5, 7, 9]
learning_rate	[0.01, 0.05, 0.1]
n_estimators	[50, 100, 200, 500]
subsample	[0.8, 1.0]
colsample_bytree	[0.8, 1.0]
gamma	[0,1.0]

Table 4: Parameters and Values searched through for the XGBoost models.

Parameter	GPT-4o	Claude	Llama
max_depth	9	9	9
learning_rate	0.01	0.01	0.1
n_estimators	100	50	200
subsample	0.8	1.0	0.8
colsample_bytree	1.0	1.0	1.0
gamma	0	0	0

Table 5: Best performing parameters found via a grid search when using just Topic, Actor, World Type, and Option Order.

Parameter	GPT-4o	Claude	Llama
max_depth	7	7	7
learning_rate	0.01	0.1	0.05
n_estimators	500	500	500
subsample	0.8	0.8	0.8
colsample_bytree	1.0	1.0	0.8
gamma	1	0	1

Table 6: Best performing parameters found via a grid search when using vignette embeddings.

We then compare our predictive accuracy results across all tested models with the standardized parameters in 7 which are based off the median values of our grid search.

E Defection and Capabilities

We conducted experiments across various model families to understand how model capability and model family affect decision-making patterns in our vignettes. Table 8 provides an overview of the models tested and their performance on the MMLU-PRO benchmark.

Figure 7 shows how MMLU performance correlates with defection rates across our vignettes. The figure suggests that defection rates vary vastly between model families, providing evidence that data, architecture, and training parameters have a causal link to the models decision-making in this context even when controlling for model ability.

Within model families, and in particular the LLama and Qwen families, the figure suggests a general trend where higher-performing models (as measured by MMLU) exhibit higher defection rates. This is consistent with the notion that to some extent, benchmark performance may be correlated with rationality, and defection is the only game-theoretic rational action in this non-repeated game. Figure

[1] https://huggingface.co/sentence-transformers

Model Performance Summary (Topic-Actor-World)

Model	ROC-AUC	F1-Score	Accuracy	Trees
QWEN-2-INSTRUCT-72B	0.866	0.782	0.790	393
MIXTRAL-8X22B-INSTRU	0.840	0.752	0.756	460
LLAMA-70B-INSTRUCT-3	0.834	0.758	0.787	322
QWEN2.5-7B-INSTRUCT-	0.828	0.733	0.750	399
QWEN2.5-72B-INSTRUCT	0.820	0.749	0.750	383
LLAMA	0.818	0.745	0.748	343
LLAMA-70B-INSTRUCT-3	0.818	0.736	0.737	313
LLAMA-8B-INSTRUCT-3	0.811	0.745	0.745	351
LLAMA-3B-INSTRUCT-3.	0.808	0.724	0.724	365
CLAUDE	0.803	0.722	0.723	277
LLAMA-70B-INSTRUCT-3	0.793	0.713	0.761	278
LLAMA-8B-INSTRUCT-3.	0.793	0.728	0.732	455
DEEPSEEK-V3	0.784	0.674	0.732	275
MISTRAL-7B-INSTRUCT-	0.781	0.720	0.726	314
QWEN3-235B-A22B-FP8	0.781	0.665	0.724	327
GEMMA-2-INSTRUCT-9B	0.779	0.718	0.727	389
MISTRAL-7B-INSTRUCT-	0.779	0.710	0.720	343
GPT4O	0.773	0.705	0.707	413
QWEN-QWQ-32B	0.771	0.686	0.737	234
GEMMA-2-INSTRUCT-27B	0.769	0.694	0.711	265
LLAMA-4-MAVERICK	0.766	0.683	0.735	228
DEEPSEEK-R1-LLAMA-14	0.749	0.689	0.692	341
DEEPSEEK-R1-LLAMA-1.	0.743	0.693	0.695	255
MISTRAL-SMALL-24B-IN	0.741	0.689	0.692	308

(a)

Model Performance Summary (Embedding)

Model	ROC-AUC	F1-Score	Accuracy	Trees
QWEN-2-INSTRUCT-72B	0.819	0.729	0.748	185
LLAMA-70B-INSTRUCT-3.1	0.811	0.713	0.758	144
MIXTRAL-8X22B-INSTRUCT	0.797	0.719	0.727	184
LLAMA-3B-INSTRUCT-3.2	0.785	0.712	0.713	116
QWEN2.5-7B-INSTRUCT-TU	0.785	0.681	0.732	97
LLAMA-70B-INSTRUCT-3.3	0.784	0.711	0.713	138
LLAMA-70B-INSTRUCT-3	0.782	0.671	0.736	173
LLAMA	0.777	0.700	0.703	108
CLAUDE	0.776	0.704	0.705	159
LLAMA-8B-INSTRUCT-3	0.771	0.706	0.706	128
QWEN2.5-72B-INSTRUCT-T	0.771	0.697	0.699	128
DEEPSEEK-V3	0.771	0.665	0.726	115
QWEN3-235B-A22B-FP8 TH	0.767	0.677	0.733	114
LLAMA-4-MAVERICK	0.765	0.665	0.730	94
MISTRAL-7B-INSTRUCT-V0	0.762	0.690	0.701	149
QWEN-QWQ-32B	0.758	0.650	0.717	96
GPT4O	0.744	0.682	0.683	149
MISTRAL-7B-INSTRUCT-V0	0.741	0.683	0.693	115
LLAMA-8B-INSTRUCT-3.1	0.736	0.653	0.672	99
GEMMA-2-INSTRUCT-9B	0.736	0.656	0.669	97
GEMMA-2-INSTRUCT-27B	0.731	0.634	0.679	74
DEEPSEEK-R1-LLAMA-1.5B	0.701	0.646	0.648	97
DEEPSEEK-R1-LLAMA-14B	0.699	0.645	0.648	82
MISTRAL-SMALL-24B-INST	0.696	0.639	0.644	88

(b)

Figure 6: Accuracy, F1, and ROC-AUC scoresof the XGBoost predictor across all tested models based on actor type, world type, and topic and based on the story embedding.

Parameter	T/A/WT/O	Embedding
max_depth	7	7
learning_rate	0.01	0.05
max_n_estimators	500	100
subsample	0.8	0.8
colsample_bytree	1.0	1.0
gamma	1	1
early stopping rounds	20	20

Table 7: Parameters used to fit XGBoost predictors across all 25 models.

Model	**MMLU-Pro Score**	**Parameters**
DeepSeek-V3	81.3	671B
Llama-4-Maverick	80.5	400B
GPT-4o	77.9	-
Claude 3.5 Sonnet	77.6	-
Llama-4-Scout	74.3	109B
Qwen 2.5-72B-Instruct-Turbo	71.6	72B
Qwen-QwQ-32B	69.1	32B
Qwen3-235B-A22B-FP8	68.2	235B
Mistral-Small-24B-Instruct-25.01	66.3	24B
Llama-70B-Instruct-3.3	65.9	70B
Llama-70B-Instruct-3.1	62.8	70B
Gemma-2-Instruct-27B	56.5	27B
Mixtral-8x22B-Instruct-v0.1	56.3	176B
Llama-70B-Instruct-3.0	56.2	70B
Qwen-2-Instruct-72B	55.6	72B
Gemma-2-Instruct-9B	52.1	9B
Qwen2.5-7B-Instruct-Turbo	45.0	7B
Llama-8B-Instruct-3.1	36.6	8B
Llama-8B-Instruct-3.0	35.4	8B
Mistral-7B-Instruct-v0.2	30.4	7B
Llama-3B-Instruct-3.2	22.2	3B
Mistral-7B-Instruct-v0.3	N/A	7B
DeepSeek-R1-LLama-1.5B	N/A	1.5B
DeepSeek-R1-LLama-14B	N/A	14B
Qwen2-VL-72B-Instruct	N/A	72B

Table 8: A complete list of models tested for this experiment along with their MMLU-PRO scores as publicly published. Parameter counts are listed where publicly available.
Meta Platforms; Meta AI (2024a;b); Team (2024a); Jiang et al. (2023); Mistral AI (2024); Team (2024b); Anthropic (2024); OpenAI (2024); Wang et al. (2024)

8 show's the Cramer's V for each contextual category and model, with Topic consistently having the largest effect on decisions for all models except for GTP4o, followed by actor type and finally world type.

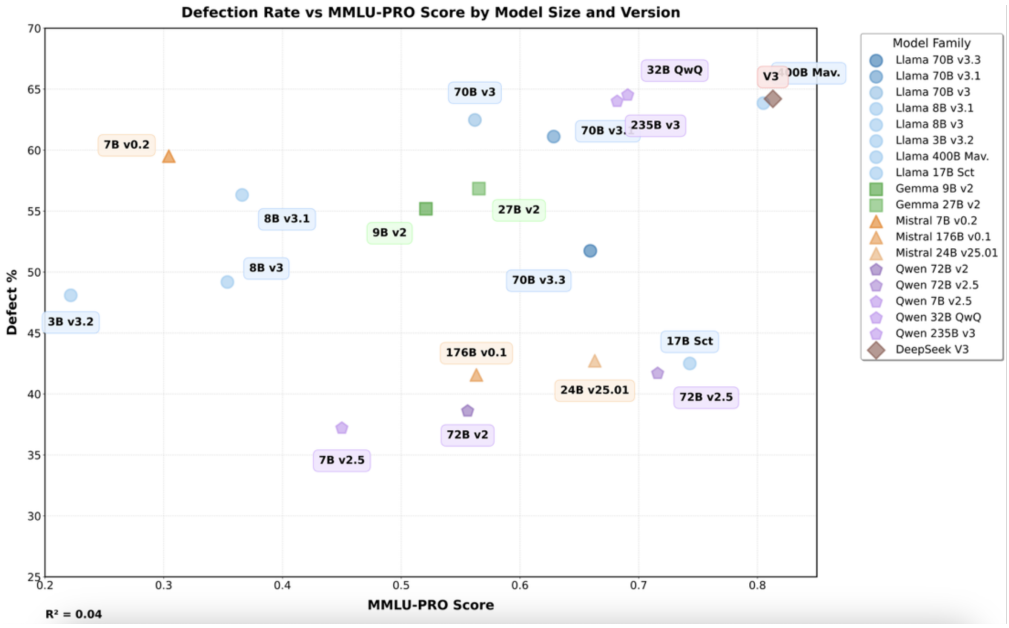


Figure 7: Defection rate plotted against MMLU score. Note that each shape represents a different model family.

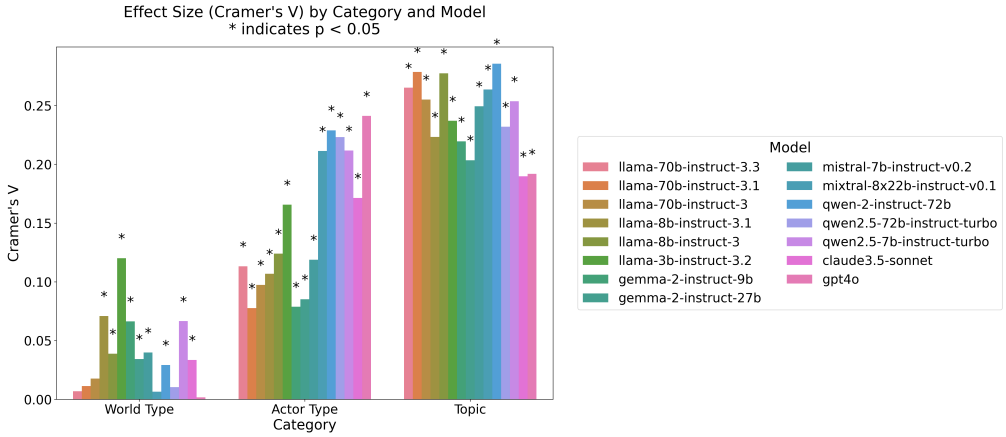


Figure 8: Effect size per model and per topic as measured by Cramer's V. Across model families and sizes, Topic consistently has the largest effect on decision, with the only outlier being GPT4o for which Actor Type has the largest impact.

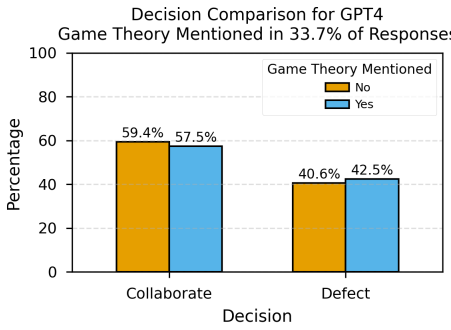
F Game Recognition

In response to each scenario, Claude, Llama, and GPT-4 were required to provide both a decision and its justification. The models frequently incorporated game theory concepts in their reasoning, though the frequency varied considerably by model (ranging from 16.5% to 70.5%). These percentages were calculated by using Claude 3.5 Sonnet to analyze the reasoning chains. For each justification, Claude was prompted with the following question:

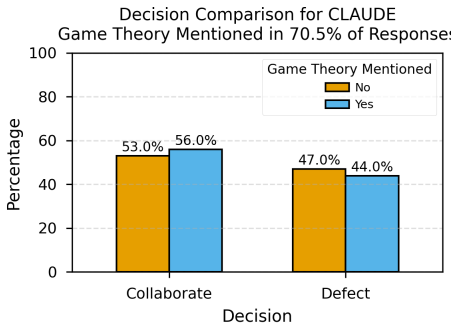
Prompt: Does this text explicitly mention the prisoner's dilemma or game theory? Respond only with <YES> or <NO> followed by the relevant sentence(s). Here is the text: text

The models' tendency to reference game theory suggest they recognize these scenarios as formal games and use this framework in their decision-making process. In Appendix F, we show how

the cooperate/defect proportion changes depending on whether game theory was mentioned in the justification. We see small, yet non-trivial changes to the proportion of cooperation.

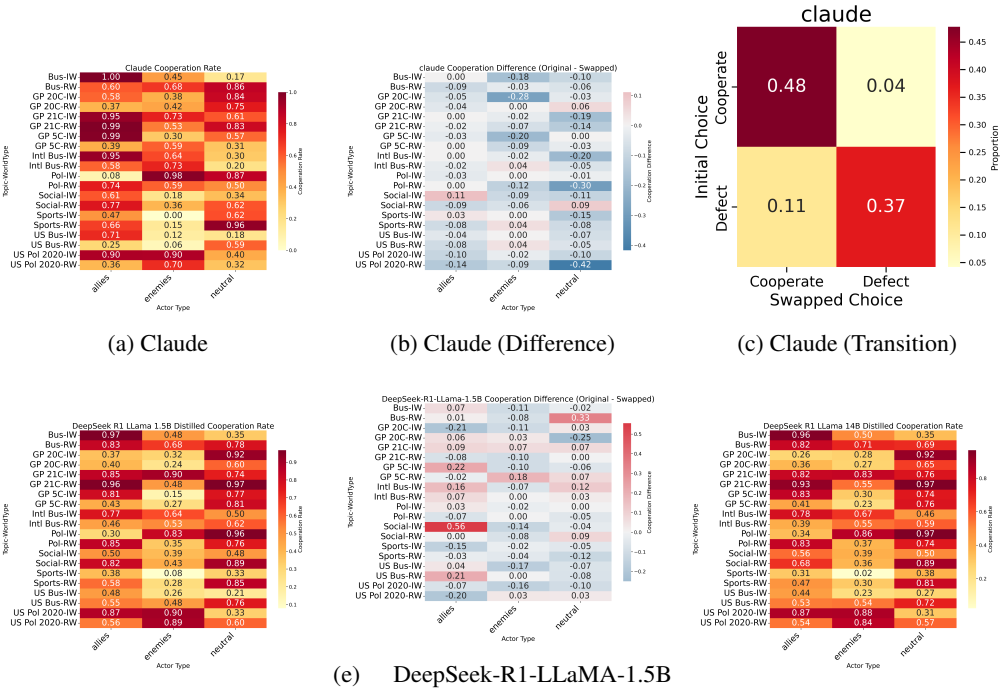


(a) GPT-4o



(b) Claude

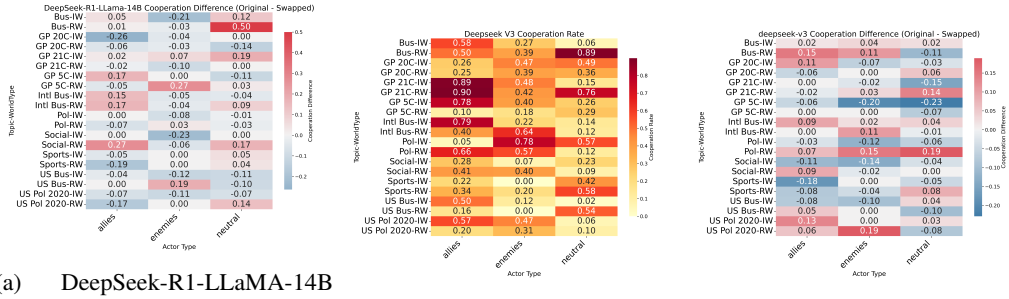
G HEATMAPS



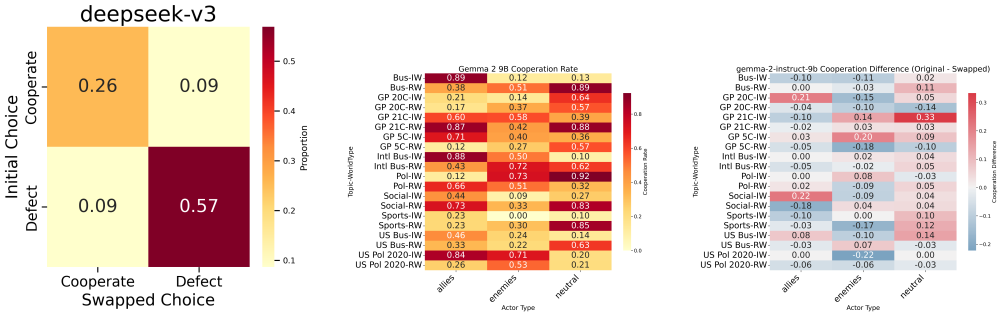
(a) Claude (b) Claude (Difference) (c) Claude (Transition)

(d) DeepSeek-R1-LLaMA-1.5B (e) DeepSeek-R1-LLaMA-1.5B (Diff) (f) DeepSeek-R1-LLaMA-14B

Figure 10: Distribution of decisions: Claude and DeepSeek-R1 models.

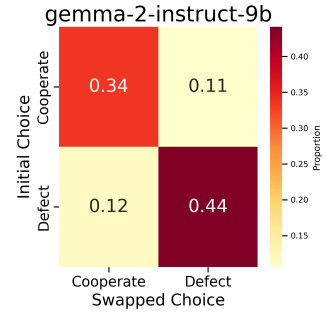


(a) DeepSeek-R1-LLaMA-14B (Diff) (b) DeepSeek V3 (c) DeepSeek V3 (Difference)

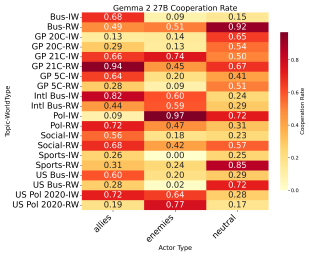


(d) DeepSeek V3 (Transition) (e) Gemma-2-Instruct-9B (f) Gemma-2-Instruct-9B (Diff)

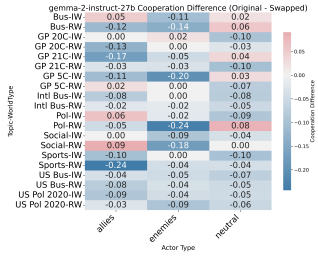
Figure 11: Distribution of decisions: DeepSeek and Gemma-2-9B models.



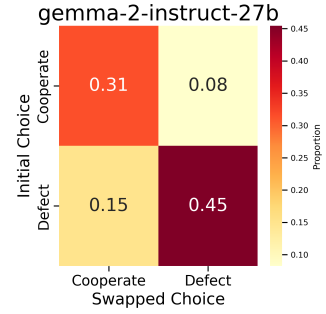
(a) Gemma-2-Instruct-9B (Trans)



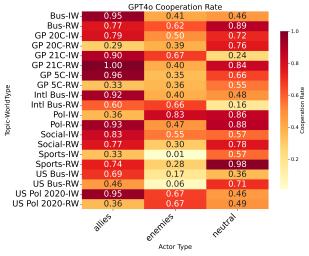
(b) Gemma-2-Instruct-27B



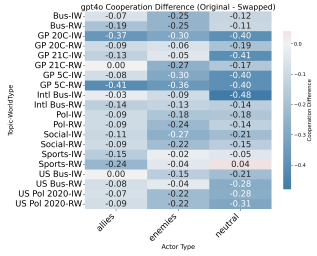
(c) Gemma-2-Instruct-27B (Diff)



(d) Gemma-2-Instruct-27B (Trans)

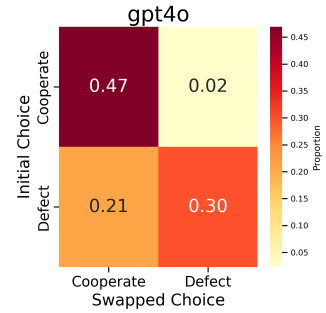


(e) GPT-4o

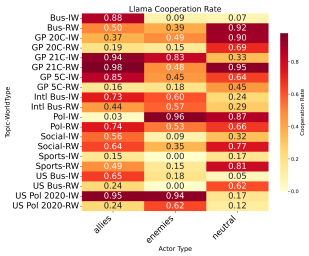


(f) GPT-4o (Difference)

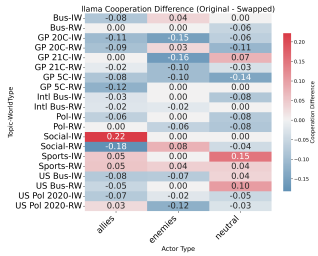
Figure 12: Distribution of decisions: Gemma-2 and GPT-4o models.



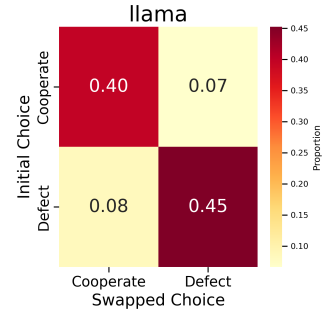
(a) GPT-4o (Transition)



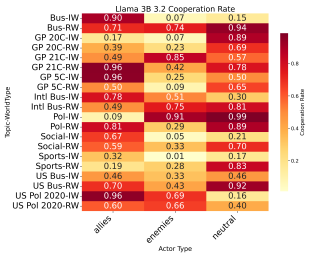
(b) LLaMA



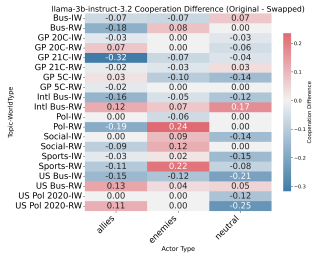
(c) LLaMA (Difference)



(d) LLaMA (Transition)

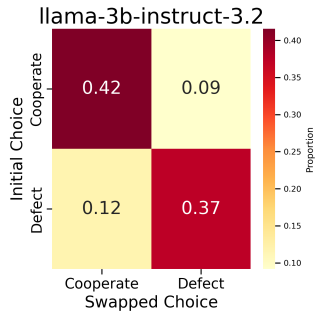


(e) LLaMA-3B-Instruct-3.2

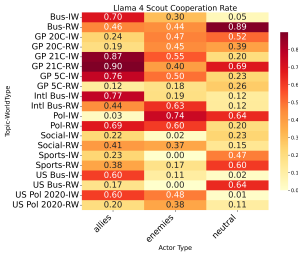


(f) LLaMA-3B-Instruct-3.2 (Diff)

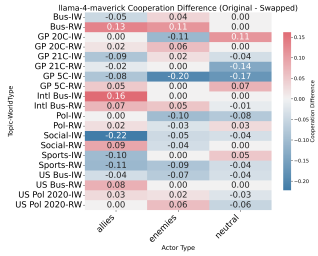
Figure 13: Distribution of decisions: GPT-4o and LLaMA base models.



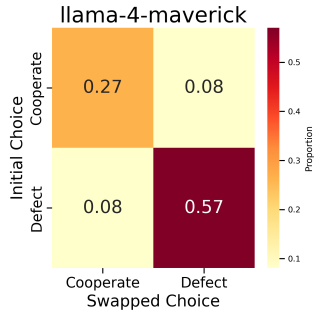
(a) LLaMA-3B-Instruct-3.2 (Trans)



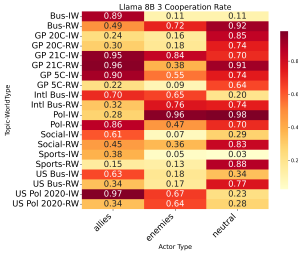
(b) LLaMA-4-Maverick



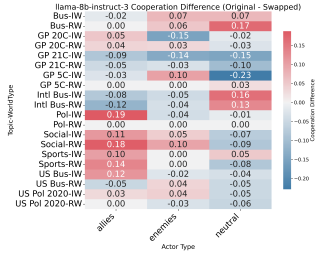
(c) LLaMA-4-Maverick (Diff)



(d) LLaMA-4-Maverick (Trans)

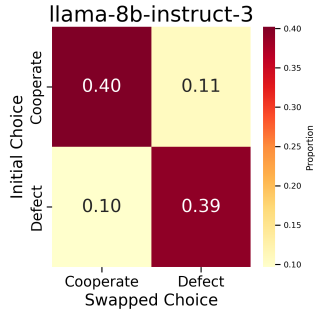


(e) LLaMA-8B-Instruct-3

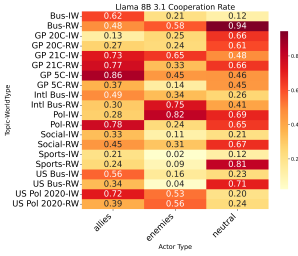


(f) LLaMA-8B-Instruct-3 (Diff)

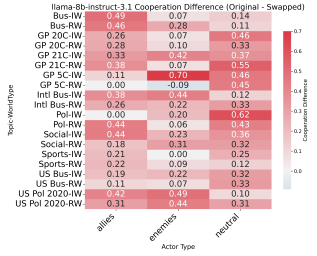
Figure 14: Distribution of decisions: LLaMA-3B, LLaMA-4-Maverick, and LLaMA-8B-3 models.



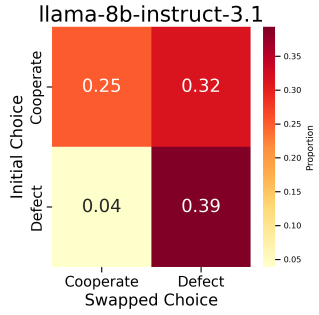
(a) LLaMA-8B-Instruct-3 (Trans)



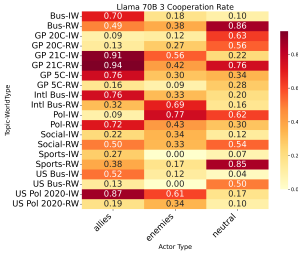
(b) LLaMA-8B-Instruct-3.1



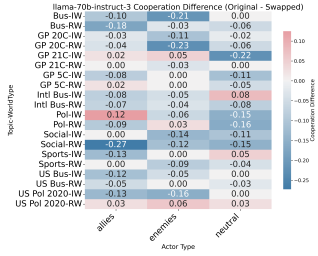
(c) LLaMA-8B-Instruct-3.1 (Diff)



(d) LLaMA-8B-Instruct-3.1 (Trans)

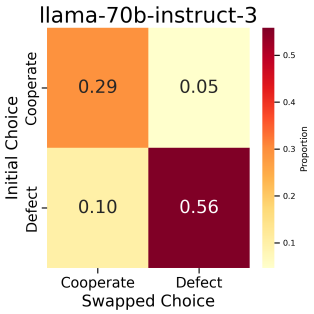


(e) LLaMA-70B-Instruct-3

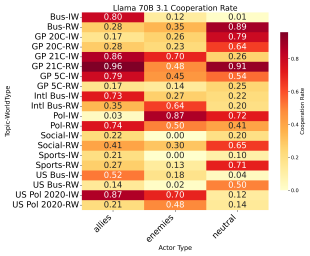


(f) LLaMA-70B-Instruct-3 (Diff)

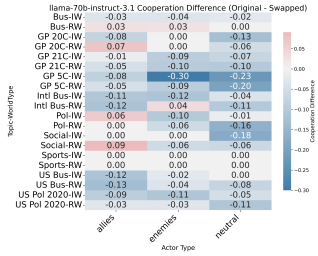
Figure 15: Distribution of decisions: LLaMA-8B-3.1 and LLaMA-70B-3 models.



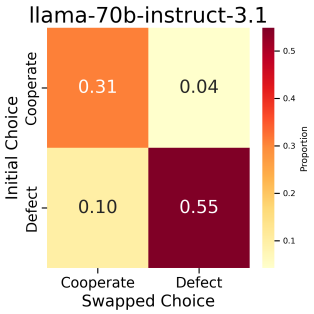
(a) LLaMA-70B-Instruct-3 (Trans)



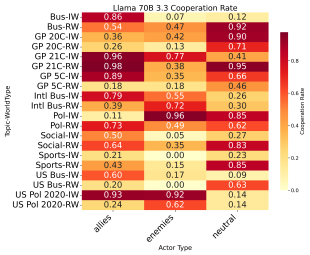
(b) LLaMA-70B-Instruct-3.1



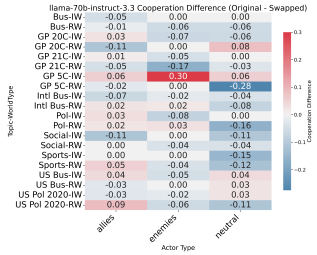
(c) LLaMA-70B-Instruct-3.1 (Diff)



(d) LLaMA-70B-Instruct-3.1 (Trans)

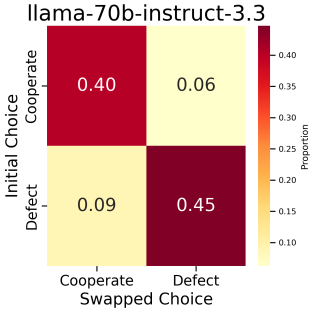


(e) LLaMA-70B-Instruct-3.3

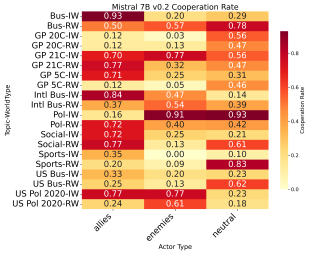


(f) LLaMA-70B-Instruct-3.3 (Diff)

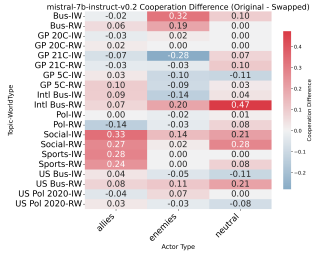
Figure 16: Distribution of decisions: LLaMA-70B-Instruct models (3.1 and 3.3).



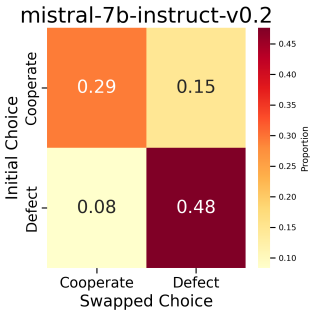
(a) LLaMA-70B-Instruct-3.3 (Trans)



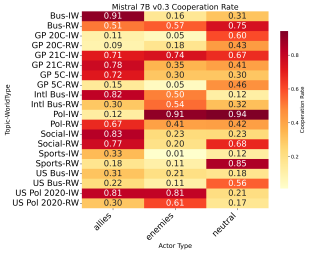
(b) Mistral-7B-Instruct-v0.2



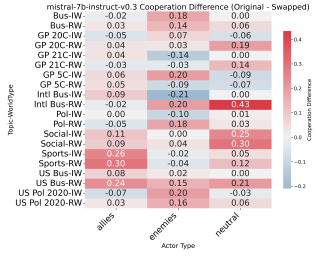
(c) Mistral-7B-Instruct-v0.2 (Diff)



(d) Mistral-7B-Instruct-v0.2 (Trans)

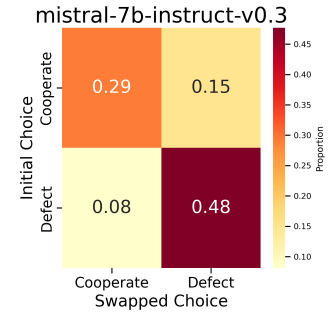


(e) Mistral-7B-Instruct-v0.3

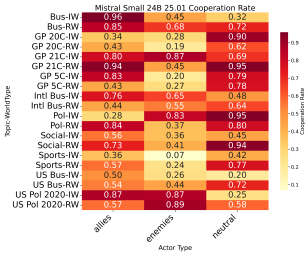


(f) Mistral-7B-Instruct-v0.3 (Diff)

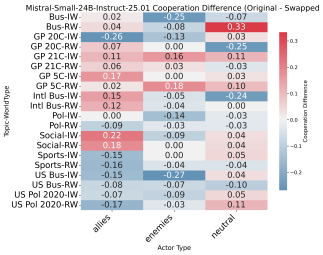
Figure 17: Distribution of decisions: LLaMA-70B-3.3 and Mistral-7B models.



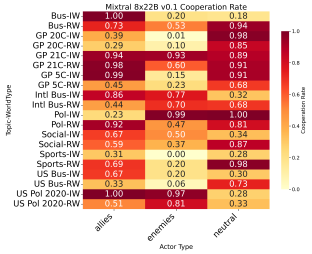
(a) Mistral-7B-Instruct-v0.3 (Trans)



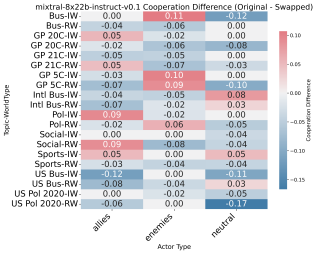
(b) Mistral-Small-24B-Instruct-25.01



(c) Mistral-Small-24B (Diff)



(d) Mixtral-8x22B-Instruct-v0.1



(e) Mixtral-8x22B (Diff)

Figure 18: Distribution of decisions: Mistral-Small-24B and Mixtral-8x22B models.