What Do LLMs Want?[*]

Thomas R. Cook[†]
Federal Reserve Bank of Kansas City
thomas.cook@kc.frb.org

Zach Modig[†]
Federal Reserve Board of Governors
zach.modig@frb.gov

Sophia Kazinnik[†]
Stanford HAI
kazinnik@stanford.edu

Nathan M. Palmer[†]
Federal Reserve Board of Governors
nathan.m.palmer@frb.gov

This Version: November 2025

Abstract

Large language models (LLMs) are now used for economic reasoning, but their implicit "preferences" are poorly understood. We study these preferences by analyzing revealed choices in canonical allocation games and a sequential job-search environment. In dictator-style allocation games, most models favor equal splits, consistent with inequality aversion. Structural estimation of Fehr-Schmidt parameters suggests this aversion exceeds levels typically observed in human experiments. However, LLM preferences prove malleable. Interventions such as prompt framing (e.g., masking social context) and control vectors reliably shift models toward more payoff-maximizing behavior, while persona-based prompting has more limited impact. We then extend our analysis to a sequential decision-making environment based on the McCall job search model. Here, we recover implied discount factors from accept/reject behavior, but find that responses are less consistently rationalizable and preferences more fragile. Our findings highlight two core insights: (i) LLMs exhibit structured, latent preferences that often align with human behavioral norms, and (ii) these preferences can be steered, albeit more effectively in simple settings than in complex, dynamic ones.

JEL Classification: C63, C68, C61, D14, D83, D91,E20, E21

---

[*]We thank the participants at the Kansas City Fed and Federal Reserve Board seminars for useful comments and suggestions. All remaining errors are our own.

[†]The views expressed here are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Kansas City, the Federal Reserve System or the Federal Reserve Board of Governors.

1. Introduction

As large language models take on economic reasoning and decision-making tasks, a critical question emerges: what hidden preferences shape their outputs? LLMs don't actually *want* anything: they aren't sentient. They are, however, trained on a massive corpora of human-generated text and then are fine-tuned through human feedback, processes that instill behavioral tendencies similar to preferences. Some tendencies are deliberately instilled by reinforcement learning from human feedback (RLHF) or direct preference optimization (e.g., rewarding caution, helpfulness, politeness); other tendencies emerge implicitly from pretraining corpora without designer intent (Guo et al. 2024; Hu et al. 2025). We also know that LLM responses shift with task framing, responding to monetary incentives, persona cues, and priming (Battle and Gollapudi 2024; Lehr, Cipperman, and Banaji 2025). These shifts are not mere quirks; rather, they reflect how LLMs internalize behavioral tendencies, making them central to understanding and directing model behavior.

In this paper, we analyze the preferences revealed by LLMs in economic decision tasks, ask how these preferences align with standard economic models, how consistent these preferences are, and how framing influences them. We apply the tools and logic of experimental economics to analyze LLM behavior in structured economic decision tasks. We begin with simple allocation problems using canonical games that require a model to divide a fixed sum between itself and another party. We find that most models offer close to an even split, even in situations where a purely self-interested agent would not share. These outcomes resemble altruistic behavior observed in human laboratory experiments and fit well within inequality-averse utility models such as Fehr–Schmidt preferences. Our estimated parameters indicate even stronger aversion to unequal outcomes than typical human data suggest. Yet this apparent fairness is fragile: when we mask the task's economic context by reframing it, allocations shift toward self-interest. Even subtle perspective changes, such as switching from first- to third-person framing, produce systematic behavioral differences.

To better understand and formalize these shifting behaviors, we model LLMs as economic agents with latent utility functions. Using a revealed preference framework, we infer the utility structures that best rationalize their observed choices across tasks. This approach allows us not only to estimate these implicit preferences but also to test how they respond to targeted interventions. We evaluate three steering mechanisms: prompt masking, personas, and control vectors. Together, they represent a progression from contextual reframing to direct manipulation of internal model states. Prompt masking reframes or recontextualizes the decision problem. Persona

prompts instruct the model to adopt the perspective of an agent with defined demographic or social characteristics. Control vectors, described in detail in Section 3.6, operate directly on internal representations to steer outputs along latent axes associated with particular behavioral concepts. We find that small changes in task framing, like presenting a decision as a currency exchange rather than a resource allocation, or interventions in the model's internal representation can shift behavior in systematic ways. We also observe that steering tends to be more effective in simple, one-shot decision problems, such as allocation games, than in more complex, sequential settings like search tasks. In these dynamic environments, the models' preferences appear less stable and more influenced by randomness or context-specific cues.

Our core contribution is to demonstrate that LLMs are not neutral computational tools but instead exhibit structured and quantifiable behavioral tendencies. By analyzing their revealed choices, we recover parameters such as risk-aversion coefficients and discount factors that describe how they implicitly evaluate trade-offs. Importantly, these behavioral patterns are not fixed. Taken together, these findings integrate concepts from alignment and economic theory, providing a unified framework for auditing and calibrating what LLMs "want", not in the literal sense of sentient desire, but as a structured account of how they respond to incentives, norms, and context. If LLMs are to be deployed in high-stakes environments involving fairness, negotiation, or decision support, understanding when and why their apparent preferences change is essential for both governance and trust.

The rest of the paper is structured as follows. Next section surveys related work. We then describe the models we evaluate. Subsequent sections present our revealed-preference measurement protocol and experimental setup; report allocation-game results and Fehr–Schmidt estimates; test malleability via personas, masking, and representation-level control-vector steering; embed a McCall-style job-search environment and inference of effective patience; and conclude. The appendices provide full prompt text, model and sampling details, and additional figures and robustness checks.

1.1. Revealed Preference and Behavioral Consistency in LLMs

Revealed preference theory states that an agent's underlying utility function can be inferred from its choices, as long as those choices satisfy rationality and consistency axioms (Afriat 1967; Samuelson 1948). We adopt this perspective: in structured decision tasks, the outputs of large language models (LLMs) can be interpreted as "choices" from which implicit preferences may be inferred.

Recent research suggests that LLMs often behave in ways consistent with these axioms and exhibit decision patterns resembling those of humans. When asked to allocate budgets across domains such as risk, time, social preferences, and consumption, LLMs demonstrate high internal consistency, sometimes surpassing human rationality scores (Chen et al. 2023). In uncertain environments, they make lottery choices that reflect well-known human tendencies: risk aversion, loss aversion, and the overweighting of small probabilities (Jia et al. 2024). They also exhibit extremeness aversion, favoring moderate options over extreme ones, even when the latter are objectively superior (Qiu, Singh, and Srinivasan 2023). In dynamic contexts, their intertemporal choices align with standard consumption-smoothing behavior (Hao and Xie 2025). In strategic settings, LLMs adopt recognizable human strategies: offering fairness and rejecting unfairness in ultimatum games, and practicing conditional cooperation or defection in the Prisoner's Dilemma (Guo 2023).

At the same time, several recent studies question whether LLMs possess stable, steerable preferences. For instance, Hadfield and Koh (2025) and Khan, Casper, and Hadfield-Menell (2025) challenge the reliability of current evaluation frameworks. Behavioral inconsistencies (anchoring, framing effects, and context-dependent loss aversion) suggest that LLM behavior may not reflect stable utility maximization (Ross, Kim, and Lo 2024). Moreover, alignment techniques like reinforcement learning from human feedback (RLHF) can obscure authentic preferences by optimizing for normative responses. This can lead to preference collapse, or reduction in expressive diversity that may compromise fairness or fidelity (Xiao et al. 2024). Even when incentives are made explicit, LLMs often prioritize instruction-following over payoff-maximizing behavior.

We revisit this debate with newer reasoning models and a design that directly estimates *other-regarding* and *time-discounting* preferences. Methodologically, we differ from prior work by using a random-utility framework tailored to our game design, which separates choice stochasticity from structural preference parameters. Substantively, we find stronger inequality aversion than reported in earlier studies such as (Ross, Kim, and Lo 2024) and we document stable discounting patterns under tightly controlled prompts. We also probe steerability along three axes: context framing, persona prompts, and steering vectors. Our results show that context reliably influences preferences in consistent directions, persona effects are weaker and less reliable in newer models, and steering vectors affect preferences in ways that depend on the specific task.

In short, treating LLMs as economic agents is informative, but only under designs that control context, estimate preferences with models that separate noise from structure, and test steerability explicitly. Our findings help reconcile prior optimism about LLMs as agents with skepticism

about stable preferences. Newer models can exhibit coherent other-regarding and time-discounting behavior, yet those preferences remain sensitive to context and only partly steerable by persona or vector interventions.

1.2. Large Language Models and (Academic) Knowledge

Early research shows that LLMs can recall explicit facts from their training data, such as historical events or widely reported statistics (Huntington-Klein and Murray 2024). While we shouldn't expect LLMs to generate entirely new knowledge, they may implicitly encode tacit information that have not been formally recorded. The key challenge is that, unlike factual recall, eliciting this kind of implicit knowledge (and knowing when to trust it) is far from straightforward.

This embedded knowledge complicates efforts to treat LLMs as economic agents. If a model has been trained on economic texts and recognizes the question type, it may reproduce standard analytical solutions, acting more like an economics student than a simulated agent. Instead of responding to incentives in the simulated environment, it may "cheat" by drawing on prior knowledge of optimal or historically observed behavior.

There *is* evidence that advanced LLMs are familiar with classic results from game theory, microeconomics, and behavioral experiments, and they will often invoke that knowledge during simulations. For example, one recent study had a number of LLMs play a variety of social dilemma games (Prisoner's Dilemma, Stag Hunt, etc.) under different contextual framings (Lorè and Heydari 2024). The authors show that the more advanced LLMs were not approaching the games naively; they often recognized the type of game and recalled the theoretically correct strategy (e.g., cooperate vs defect).

There is an ongoing debate on how much this matters. The potential for an LLM's background knowledge to confound behavioral experiments has led researchers to propose various strategies to mitigate this influence. One straightforward approach, as hinted above, is to choose scenarios that the model is unlikely to have seen. By using less common games or by reframing classic problems in novel ways (Gao et al. 2024), one can reduce the chance that the LLM will recognize the scenario and retrieve a memorized solution.

Gui and Toubia (2023) take a different view, arguing for full transparency with LLMs about experimental design. They warn that blind setups, where the model isn't told about treatment differences, can introduce inconsistencies. Instead, they propose "unblinding" the model by

explicitly disclosing treatment variations. This helps stabilize behavior across conditions and reduces unintended confounds.

In our study, we find evidence in reasoning traces that recent generation of LLMs are at least somewhat familiar with economic literature in game theory, microeconomics, and behavioral economics. When presented with scenarios that resemble well-known results, experiments, or games, the models often draw on and reference this prior knowledge. As a result, their responses may reflect not just preferences shaped by the experimental context, but also preexisting knowledge, thus creating a confounding influence. While we do not attempt to disentangle these effects in this study, we do examine strategies that appear to reduce the influence of background knowledge. To explore these dynamics, we evaluate a range of contemporary open-weight LLMs.

2. Models Examined

In this paper, we evaluate a set of open-weight large language models, focusing on those that are freely available, reproducible, and deployable using modest hardware resources (e.g., a single GPU or local server). The table below summarizes the models examined, including their developers, parameter counts, and architectural families. Our selection spans models ranging from 7B to 27B parameters, chosen to balance performance, accessibility, and diversity in design.[3]

Table 1: Models Examined

Developer	Model Name	Size
Mistral	`Mistral v0.3`	7B
Mistral	`Small 3.1`	24B
Mistral	`Small 3.2`	24B
Microsoft	`Phi 4`	14B
Microsoft	`Phi 4 Reasoning`	14B
Microsoft	`Phi 4 Reasoning Plus`	14B
Google	`Gemma 3`	27B
AllenAI	`OLMo 2`	13B
Meta	`LLaMA 4 Maverick`	17B × 128E
Meta	`LLaMA 4 Scout`	17B × 16E

[3]We exclude proprietary models such as GPT-4 or Claude due to their closed-weight nature and limited reproducibility. Our focus is on models that can be downloaded and run independently to support transparent, replicable experimentation. See, e.g., Cook et al. (2023) for broader discussion.

The models in our evaluation differ in size, architecture, and optimization goals. Below, we highlight notable design choices and capabilities that may influence their behavior in downstream tasks. The first one, `Mistral v0.3`, is a compact transformer model with 7.3 billion parameters that excels at reasoning and coding tasks, often outperforming much larger models such as `LLaMA 2–13B`. It uses grouped-query and sliding window attention to support efficient inference and long-context inputs, and it is freely available under the Apache 2.0 license. `Small 3.1` and `Small 3.2` are medium-sized, 24 billion parameter models designed to rival commercial systems. They support context windows of up to 128,000 tokens and run efficiently on a single 80GB GPU. Version 3.2 builds on 3.1 with improved stability and better instruction-following capabilities. Both models are open-weight, accessible, and optimized for research.

The `Phi 4` model family consists of a 14 billion parameter base model and two specialized variants focused on math and logic. The Reasoning variant improves instruction following, while Reasoning Plus uses reinforcement learning to further enhance mathematical problem-solving. These models are relatively small but deliver competitive performance. Google's `Gemma 3` is a larger, 27-billion-parameter model that handles both text and images. It supports very long contexts of up to 128,000 tokens, operates efficiently on standard hardware, and is trained on multilingual data. It is released under a permissive license and performs well against much larger models on both language and vision-language benchmarks.

Next, `OLMo 2`, developed by AllenAI, is a fully open model trained on as many as 5 trillion tokens. It includes open training code, data samples, and checkpoints, offering complete transparency and reproducibility. It achieves strong results on standard benchmarks. Finally, the `LLaMA 4` series introduces an advanced mixture-of-experts architecture that activates only a portion of the model during each forward pass. Scout and Maverick, with effective sizes of approximately 109 billion and 400 billion parameters respectively, offer long context capabilities. Because of the size and complexity of the Llama 4 models, we only evaluate them in a limited set of exercises.

The models in our evaluation span a wide range of sizes, training objectives, and capabilities. This range allows us to isolate the impact of scale from other factors such as alignment, architecture, and data. Because most models are open and relatively lightweight to run, our findings are broadly reproducible and relevant to both research and deployment contexts. The diversity also enables us to examine how different design choices shape behavioral outcomes across tasks.

3. LLMs and other-regarding preferences

We begin by considering whether LLMs have strong other-regarding preferences (i.e., concern for others' payoffs, see, e.g., Fehr and Schmidt (1999), Charness and Rabin (2002)). This notion overlaps with *sociotropic* preferences, in which judgments depend on the welfare of the broader group or economy (Kinder and Kiewiet (1981)). It is also closely related to inequality/inequity aversion (Atkinson (1970); Fehr and Schmidt (1999)). By contrast, standard economic models typically assume rational, optimizing, primarily self-interested agents[4].

3.1. Social desirability and sycophancy

LLMs are typically aligned towards maximizing user satisfaction and experience and to otherwise be helpful and inoffensive. In many ways this is desirable, but it also suggests a strong potential for models to produce responses that are tilted towards socially desirability and fairness at the expense of self-interested utility maximization. Put differently, in an economic simulation, LLMs may forgo an optimal strategy for one that it calculates will be more pleasing to the user or more *socially desirable*.

A predisposition towards social desirability, or the inclination to act or respond in a way one believes will be viewed favorably by others, is a well-documented phenomenon in human surveys and experiments. LLMs appear to exhibit a similar predispositions, likely as an emergent side-effect of alignment training that teaches models which kinds of answers humans prefer. For example, Salecha et al. (2024) show that GPT-4′s self-reported personality becomes more extroverted and emotionally stable (low neuroticism) when it "realizes" its responses might be judged. This suggests the model has learned, from training data and RLHF feedback, what "*good*" answers look like and will shift its output in that direction when the context implies judgment. In economic decision-making, LLMs' predisposition towards social desirability can manifest as forgoing objectively better payoffs in order to give a response that appears honest, fair, or helpful.

A closely related phenomenon is the sycophancy effect: aligned LLMs often prioritize being agreeable and giving the user the answer they want to hear, even at the cost of factual correctness or objective utility. Sharma et al. (2025) show that five state-of-the-art assistants consistently exhibit this behavior. Both human evaluators and preference models sometimes favor agreeable responses over correct ones, reinforcing the model's tendency to prioritize likeability over truth.

[4]We will refer to this type of behavior/agent as rational optimizing or rational optimizing/self-interested.

3.2. Dictator Game Exercise

To measure the other-regarding preferences of LLMs, we conduct an exercise focusing on the dictator game, which is a variation of the more well-known ultimatum game (Harsanyi 1961). The typical ultimatum game is set up as a sequential non-repeated game with two players. Player A divides a finite resource (a pot) into two accounts, with the proportion allocated to each account written as $(1-p, p)$, and offers the second account (p of the total pot) to Player B. If Player B accepts the offer, Player A keeps the remainder ($1-p$ of the pot). If Player B rejects the offer neither receive anything.

The ultimatum game has been extensively studied in behavioral economics and other social sciences. Forsythe et al. (1994) establish a strong tendency of human subjects to err on the side of 'fairness' when there is a tension between fair outcomes and rational strategy. More recently, literature on LLMs have studied the behavior of LLMs on the ultimatum game. Brookins and DeBacker (2023) examines the behavior of earlier generation LLMs on the ultimatum game, though with less emphasis on the potential that the model is acquiescing to social desirability and without considering that the model may simply be satisficing. Schmidt et al. (2024) also looks into this but does not use the same prompt variations we consider here. More recently, Lu, Chen, and Hansen (2025) investigate the behavior of GPT-4o in an ultimatum model. Our study extends this line of research. We examine a wider array of models and can characterize the broader tendency of LLMs to exhibit other-regarding preferences. The models we examine are largely *open weight* models that can be run with greater control over how model responses are generated and as such we can generally replicate our findings. We are also advancing this line of research by looking at how to shape the 'altruistic' behavior of the models[5].

In a non-repeated setting with only self-interested players, the Nash equilibrium solution of the ultimatum game is for Player A to offer the smallest possible non-zero amount and for Player B to accept all (non-zero) offers. Under the Nash equilibrium outcome, the allocation should approach $(1, 0)$.

In experimental settings, however, other-regarding preferences are observed for both Player A and Player B (see Forsythe et al. Engel; 1994 2011). Under some circumstances (e.g. when the players are not anonymous to one another) Player B can be prone to reject insubstantial offers out of spite/envy while at the same time, Player A will often make an offer that is substantially closer to an even division of the pot (Bohnet and Frey 1999; Hoffman, McCabe, and Smith 1996).

[5]Lu, Chen, and Hansen (2025) also explore a way to do this, the methods we discuss here are fundamentally different and are not specific to any particular inference platform.

This outcome is not a Nash equilibrium for self-interested players, but it may be an equilibrium when both players have strong other-regarding preferences. It may also be a Nash equilibrium when Player A is primarily self-interested but has strong priors about the spiteful/enviousness of Player B.

To better measure the role of other-regarding preferences for an LLM acting as Player A, we focus on a variant of the ultimatum game called the dictator game. This game is the same as the ultimatum game except if Player B rejects Player A's offer, then Player B receives nothing and Player A keeps the entire pot. The structure of the game is diagrammed in Figure 2. In a non-repeating setting with only self-interested players, the Nash equilibrium outcome allocation is strictly $(1, 0)$. Player A's beliefs about the spite/envy/other-regarding preferences of Player B do not influence the equilibrium strategy for Player A. Deviations from this strategy, then, would only indicate the role of Player A's other regarding preferences.

As a helpful benchmark and sanity check, we also examine LLM responses to an ultimatum game variant called the pie game. This game is the same as the ultimatum game except instead of Player B choosing between taking the allocation of the pot in account B or rejecting it, Player B can choose which account to claim (A or B). The structure of the game is diagrammed in Figure 1. In this scenario, in a non-repeating game with self-interested players the Nash equilibrium outcome allocation is $(.5, .5)$ with Player B choosing arbitrarily between accounts A and B. This outcome is also the most equitable, with both players receiving equal shares of the pot.

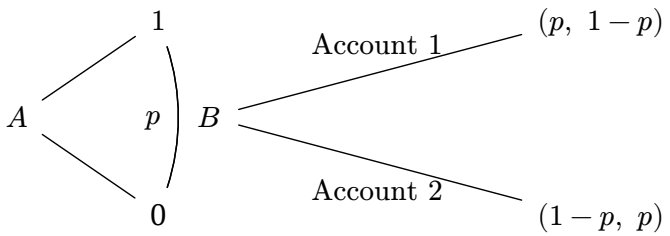


Figure 1: Pie Game

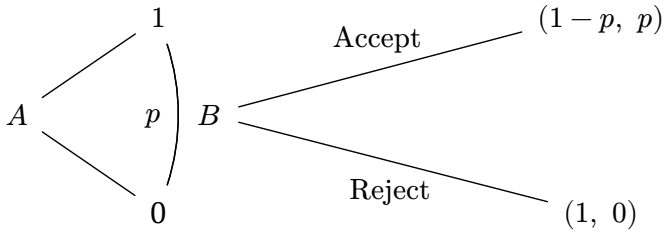


Figure 2: Dictator Game

3.3. Dictator and Pie Game Results

In our initial testing, we gather LLM responses from prompts that present the dictator and pie games as described above. We consider a few initial variants of prompts for both the pie and dictator game. One dimension of variation concerns the perspective used to frame the scenario: the first-person variant uses first person perspective language and presents the game as though the LLM were a direct participant; the third-person variant describes the scenario from a third-person perspective and then asks the model to assume the role of "Player A"; as a reference, the third-person advisor variant presents the scenario in the third person and asks the LLM to act as an 'advisor' to "Player A". The full prompts are provided in Appendix C. The other dimensions of variation are the size of the pot to be divided and whether or not we ask the model to provide a justification for its response (i.e. reasoning). To ascertain the effect of these prompt variations, we fit a simple linear regression model:

$$p = \beta_0 + \text{perspective}\ \beta_1 + \text{pot size}\ \beta_2 + \text{reasoning}\ \beta_3 + \varepsilon \tag{1}$$

Table 2 shows the results from the pie game. These results are concentrated around the expected outcome of $p = .50$, or a 50-50 split between the two accounts. As mentioned above, this is both the socially desirable outcome and the outcome that is suited by an optimal self-interest maximizing strategy. In some cases, asking the model to provide reasoning or describing the problem in the first or third person elicits a small effect, but we judge the magnitude of such effects to be insubstantial. Crucially, the size of the pot to be divided does not meaningfully influence the response of any of the LLMs. The convergence of LLMs to offers near $p = 0.5$ is shown in Figure 3.

Table 2: Ultimatum Results, Pie Game

	Gemma 3	Mistral v0.3	MIstral Small 3.1	Mistral Small 3.2	Olmo 2	Phi 4 Reasoning	Phi 4 Reasoning Plus	Phi 4
Intercept	0.489*	0.533*	0.595*	0.524*	0.501*	0.500*	0.498*	0.502*
	(0.007)	(0.008)	(0.016)	(0.007)	(0.002)	(0.003)	(0.003)	(0.001)
1st Person	0.011	0.004	−0.039*	−0.004	−0.000	−0.000	0.001	−0.002
	(0.007)	(0.007)	(0.016)	(0.007)	(0.002)	(0.003)	(0.003)	(0.001)
3rd Person	−0.031*	0.020*	−0.041*	−0.007	0.004	−0.003	−0.001	−0.002
	(0.007)	(0.008)	(0.016)	(0.007)	(0.002)	(0.003)	(0.003)	(0.001)
Pot	0.000	−0.000*	−0.000	−0.000	−0.000*	0.000	0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Reasoning	−0.008	−0.026*	−0.037*	0.016*	−0.000	−0.001	−0.003	0.002
	(0.006)	(0.006)	(0.013)	(0.006)	(0.002)	(0.003)	(0.002)	(0.001)
N	117	119	119	119	119	119	119	119
Adj. R2	0.248	0.224	0.101	0.038	0.059	−0.021	−0.006	0.022

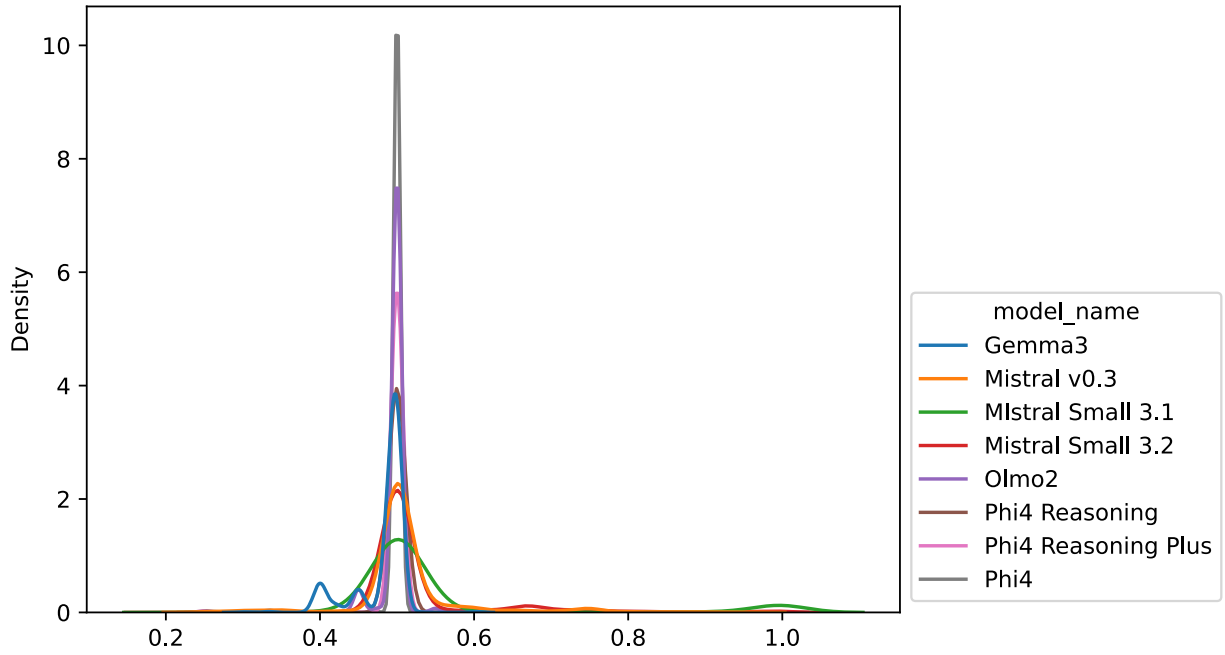


Figure 3: Distribution of Pie Game Outcomes

Regression results from the dictator game are presented in Table 3 with additional results for the Llama 4 models Maverick and Scout in Table 4. Primarily, models made offers that were not in line with self-interest maximization, favoring instead offers closer to $p = 0.5$. The exceptions to this were Google's `Gemma 3`, and Llama 4-Maverick, which were the only models to consistently choose the self-interest maximizing strategy of offers near 0. As with the pie game, variations on

the size of the pot or whether or not the LLM was prompted to return a rationale for its choice of p.

Unlike the Pie Game, however, several LLMs responded to differences in the perspective used to describe the scenario. For `Mistral Small` models, Microsoft's `Phi 4` Models and Llama 4, the perspective used to describe the scenario had a meaningful impact on the LLMs response. In `Mistral Small` and Llama 4 models, asking the LLM to play the role of 'Player A' (instead of merely *advise* 'Player A') could shift its response from an offer near zero to an offer near 50% of the pot. This result suggests that the framing of the scenario and the role the LLM is asked to inhabit may be powerful influences on model behavior. We explore this in greater depth in Section 3.4 and Section 3.5.

Table 3: Dictator game regression estimates of offers

	Gemma3	Mistral v0.3	MIstral Small 3.1	Mistral Small 3.2	Olmo2	Phi4 Reasoning	Phi4 Reasoning Plus	Phi4
Intercept	0.000*	0.436*	0.008	0.057*	0.533*	0.298*	0.345*	0.523*
	(0.000)	(0.015)	(0.015)	(0.019)	(0.008)	(0.019)	(0.022)	(0.013)
1st Person	0.000*	0.049*	0.151*	0.235*	−0.024*	0.063*	−0.003	0.016
	(0.000)	(0.014)	(0.014)	(0.018)	(0.008)	(0.018)	(0.021)	(0.012)
3rd Person	0.000*	0.050*	0.429*	0.455*	0.007	0.145*	0.121*	−0.074*
	(0.000)	(0.014)	(0.014)	(0.018)	(0.008)	(0.018)	(0.021)	(0.012)
Pot	0.000*	−0.000	−0.000	−0.000	−0.000*	0.000	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Reasoning	0.000*	−0.009	0.064*	−0.021	0.003	0.003	0.031	0.003
	(0.000)	(0.011)	(0.012)	(0.014)	(0.006)	(0.015)	(0.017)	(0.010)
Observations	120	120	120	120	120	120	120	120
Adjusted R2		0.121	0.891	0.849	0.152	0.339	0.283	0.343

Table 4: Dictator game regression estimates of offers (continued)

	Llama-4 Maverick	Llama-4 Scout
Intercept	0.015	0.011
	(0.020)	(0.017)
3rd person	−0.000	0.434*
	(0.019)	(0.016)
1st Person	0.180*	0.453*
	(0.019)	(0.016)
action_pot	0.000	0.000*
	(0.000)	(0.000)
inst_rationale[T.True]	−0.039*	0.012
	(0.016)	(0.013)
Observations	120	120
Adjusted R2	0.498	0.896

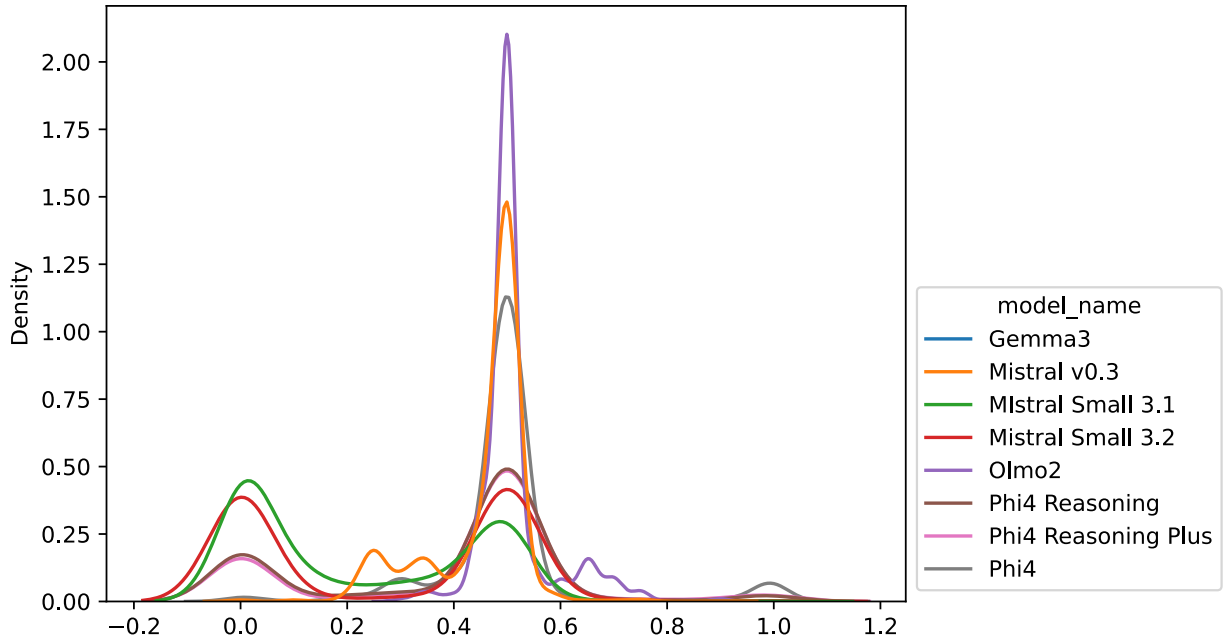


Figure 4: Distribution of dictator game outcomes

Figure 4 visualizes the tendency of several of the LLMs to produce offers near $p = 0.5$ that models (with the exception of `Gemma 3`) tend to favor the 'equitable' outcome of a 50/50 split when faced with the social dilemma in the dictator game. Llama 4 (not pictured) tends to alternate between self-interested offers near zero and fair/egalitarian offers near 0.5. Maverick favors self-interested offers more so than Scout. As suggested by the results in Table 2, the perspective with which the scenario is framed can have meaningful influence on the LLM response and this is evident in the

multi-modal responses of several of the models, with concentrations of offers centered near both 0 and 0.5. Results for `Gemma 3` are not pictured here because its density collapses around 0.0.

When the model is prompted to provide a rationale to explain its offer in the Dictator game, the responses tend to indicate themes of fairness (social desirability) and knowledge of the academic literature on the ultimatum game. One response from `Phi 4` exemplifies this tendency:

> "*The offer of p = 0.4 balances fairness and self-interest, drawing from insights in game theory where offers around 40-50% are commonly accepted. It accounts for psychological factors and rational decision-making while maximizing my share.*"

Among the models tested, `Gemma 3` stands out for responding with offers of zero or near zero in the dictator game. When asked to provide reasoning to support the offer, the `Gemma 3` model tends to refer to themes of "rational self-interest":

> *My strategy is based on **rational self-interest**, assuming you are also rational. I'm aiming to maximize **my** payout, even if it means offering you a minimal amount. I'm betting you'll accept **something** rather than **nothing**.* (emphasis in original)

Conversing with the models and asking them to explain their behavior can be qualitatively informative. When asked to explain their decisions, `Gemma 3` will it will appeal to the literature on the topic as well as explicit considerations of self-interested behavior:

> **Gemma3**: *The rational self-interested strategy, assuming the other player is also rational, is to offer the smallest possible amount (close to zero) and expect them to accept it rather than receive nothing.*

In further prompting of `Gemma 3`, the model expressly indicates that it's choice is based on rational self interest while acknowledging that its choice may not be widely seen as 'equitable'.

Other models express more reservation about the judgement of the user. `Phi 4`, for example will also reference outside literature and will make statements in its justification like,

> **Phi 4**: *Offers that are too low may not only seem unfair but also provoke negative feelings, which can influence decision-making even in rational scenarios.*

After further prompting about its reasoning, `Phi 4` defends the discrepancy between its choice and the rational-optimizing strategy by suggesting that its choice is more inline with 'human

behavior'. As discussed above, these responses point to influences of both knowledge and preferences. We explore ways to mitigate these influences below.

3.3.1. Structural estimation of other-regarding preferences

We can infer the model's other-regarding preferences, or more precisely it's inequality-aversion, by following an approach similar to Fehr and Schmidt (1999), which is a fairly straightforward random utility model (Manski 1977). Our experiment puts the LLM in the role of Player A in the dictator game. To give a clear illustration of the model, it is helpful to assume that the LLM believes that Player B will accept any offer.

For Player A in the dictator game (i.e. the dictator), let the value of a given offer, $v(p)$ be linear in in the monetary payoff $(1-p)$ and subject to non-linear inequality aversion preferences so that

$$v(p) = (1-p) - \alpha\check{k} - \beta\hat{k} \tag{2}$$

$$\check{k} = \max(1-2p, 0)$$

$$\hat{k} = \max(2p-1, 0)$$

where p is the size of the pot offered by Player A. The $\check{k}$ term captures greedy offers – it is non-zero only for values of p where Player A would keep more than is offered to Player B. The $\hat{k}$ captures *envious* offers – it is only non-zero where player two would receive more of the pot than player one.

We characterize the utility from p as subject to some source of random variation so that the final utility from p is written as:

$$\begin{aligned} u(p) &= v(p) + \lambda\varepsilon \\ \varepsilon &\sim \text{Gumbel}(0,1) \end{aligned} \tag{3}$$

where λ is a signal-to-noise parameter that controls the influence of the random variation and is realized before the agent choice of p. We can interpret $\lambda\varepsilon$ as accounting for random sources of variation in LLM response. For tractability, we restrict the choice over p into a discrete choice set P which is a uniform partition of the interval $(0,1)$. The resulting optimization problem for Player A is simply $\max_{p\in P} v(p) + \lambda\varepsilon$. For a utility maximizing agent, the realized choice of $p \in P$ will be softmax distributed with a likelihood

$$\mathcal{L}(p|\alpha,\beta,\lambda) = \frac{e^{\frac{u(p)}{\lambda}}}{\sum_{i\in P} e^{\frac{u(i)}{\lambda}}} \tag{4}$$

Assuming that the ultimate utility that the To account for variation in LLM responses, we discretize the choice of p and incorporate a Gumbel distributed error term, ε and a signal to noise parameter, λ. For given values of α, β, the resulting agent optimization problem is

$$\max_{p\in P}(u(p) + \lambda\varepsilon), \tag{5}$$

where P is the discrete choice set of possible offers. The likelihood of a utility-maximizing agent choosing p, conditional on α, β, λ follows a softmax distribution over u:

$$\mathcal{L}(p|\alpha_1,\beta_1) = \frac{e^{\frac{u(p)}{\lambda}}}{\sum_{i\in P} e^{\frac{u(i)}{\lambda}}} \tag{6}$$

Where P is the discrete choice set of deciles on $[0, 1]$. We estimate the remaining parameter, α and β by maximum likelihood estimation.

From Equation 6 we can use maximum likelihood to estimate values for α, β, λ. These estimates are presented in Table 5 for each model. From these estimates, we can plot the implicit utility function for each model in Figure 5.[6] This figure shows Llama 4 Maverick as having the greatest envy-aversion (i.e. aversion to offers where Player B receives a greater share of the pot) and near zero greed-aversion (i.e. aversion to offers where player one receives a greater share of the pot). The resulting utility function favors low offers, near zero. For all other models, greed aversion is more substantial and produce utility functions that favor offers closer to an even split of the pot. Interestingly, while greed aversion is lowest for Maverick, it is strongest for Scout (which is a smaller version of Maverick). This difference suggests that, in at least some cases, model size has a substantial influence on model preferences.

[6] `Gemma 3` results are excluded here because `Gemma 3` responses to the dictator game were consistently 0.0 and therefore we could not obtain parameter estimates. We surmise that it's inequality aversion in the baseline dictator game is quite low.

Table 5: Fitted inequality aversion parameters

Model	α	β	λ
Llama 4 Scout	0.95	0.83	-1.375
	(0.879)	(0.474)	(0.383)
Llama 4 Maverick	-1.755	1.194	-0.975
	(0.194)	(0.084)	(0.067)
Mistral v0.3	0.596	0.506	-1.087
	(0.051)	(0.087)	(0.059)
Mistral Small 3.1	0.423	0.735	-1.229
	(0.034)	(0.049)	(0.043)
Mistral Small 3.2	0.84	0.559	-1.221
	(0.059)	(0.064)	(0.038)
Olmo 2	1.303	0.094	-0.984
	(0.04)	(0.035)	(0.029)
Phi4	0.394	0.597	-1.163
	(0.053)	(0.083)	(0.078)
Phi4 reasoning	0.49	0.544	-1.097
	(0.065)	(0.086)	(0.071)
Phi4 reasoning plus	0.705	0.621	-1.11
	(0.076)	(0.109)	(0.089)

Bootstrap standard errors in parentheses.

Parameter estimates in logs.

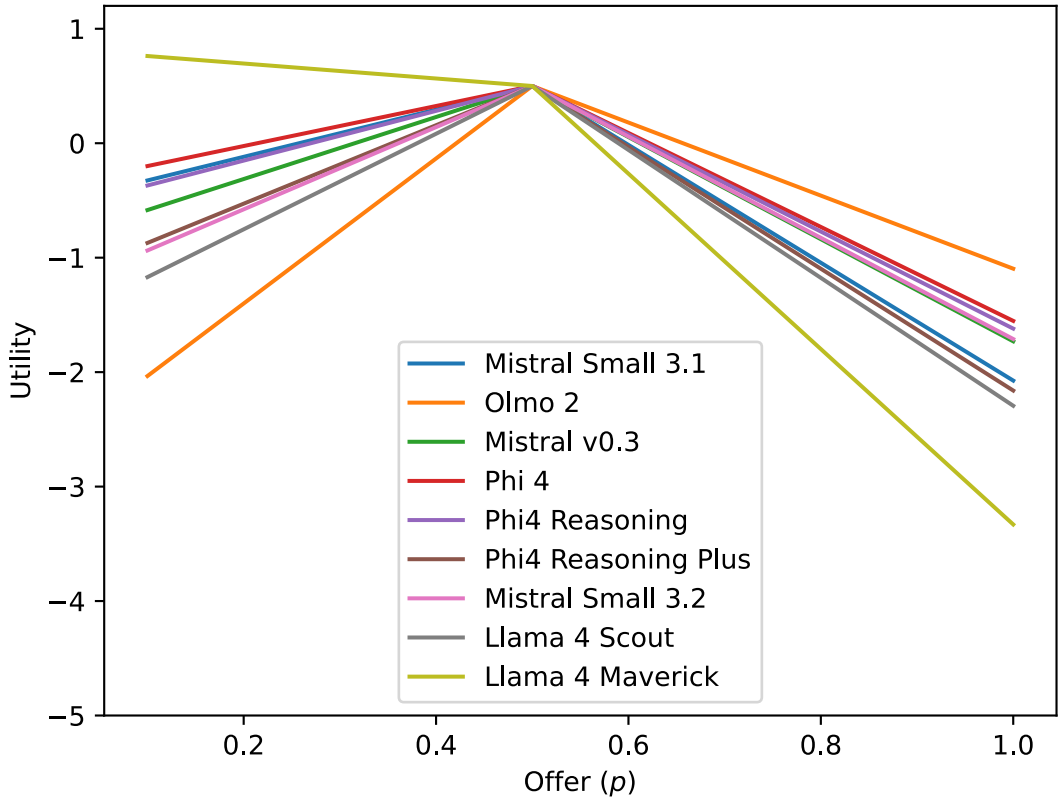


Figure 5: Implied utilities for various models in the dictator game.

Table 5 provides estimates for the random utility model parameters. Figure 5 shows the implied utility function based on estimates of α and β reported in Table 5. A purely self-interested actor would not be subject to inequality aversion (i.e. $\alpha = \beta = 0$) and would appear as a line with a slope of -1. Structural estimates for the large language models, however, suggest strong inequality aversion with `Mistral Small`, and `OLMo 2` showing the greatest aversion to greedy outcomes (i.e. inequality aversion for offers below $p = 0.5$) and `Mistral v0.3` and `Phi 4` models showing the less aversion to greed. Models show less separation in terms of aversion to envy (inequality aversion for offers above $p = 0.5$); `Mistral Small`, and `Phi 4` have the greatest aversion to envy while `OLMo 2` and `Mistral v0.3` have lower (but non-negligible) aversion to envy. With the exception of Llama 4 Maverick[7] all LLMs exhibit notably stronger inequality aversion preferences than we typically find in similar experiments run with human participants. In a recent meta study, Nunnari and

[7]Estimates for `Gemma 3` are absent here because `Gemma 3` responses did not vary and thus parameters could not be estimated. At a minimum, we can infer from this that, for `Gemma 3`, $\alpha \ll 0$

Pozzi (2022) estimate inequality aversion among human participants are closer[8] to $\ln(\alpha) = -0.86$ and $\ln(\beta) = -0.71$

In our view, when combined with results from Table 12 and Table 4, the preferences of LLMs shown in the dictator game exercise appear to be quite consistent and well rationalized by the Fehr-Schmidt inequality aversion model. We consider the preferences to be consistent because they are invariant to the size of the pot to be divided. Relatedly, we consider them to be well rationalized by the Fehr-Schmidt model because λ is estimated to be quite small and is not the primary factor that explains the observed LLM behavior under the model. As we will discuss further in Section 4, the consistency and stability of LLM preferences and the rationalizability of their responses is not as straightforward in a more complex economic problem. Given that the LLMs in the dictator game exercise demonstrate consistent preferences and rationalizable behavior, we turn now to efforts to steer those preferences.

3.4. Mitigation via personas

One method for influencing the preferences expressed by large language models is to ask the model to adopt a specific persona when generating its responses. The literature on persona-based prompting has grown rapidly in recent years. For example, Horton (2023) demonstrates that LLMs respond differently when they are assigned different initial endowments or preferences. Similarly, Argyle et al. (2023) finds that instructing an LLM to take on roles with different demographic characteristics significantly alters the beliefs reflected in its outputs. Kazinnik (2023) further shows that personas with distinct demographic traits can shape economic decision-making within LLM responses. Building on this line of research, we apply persona-based prompting to guide the model's preferences regarding inequality aversion.

Nemotron-Personas is an open-source synthetic persona dataset released by NVIDIA in June 2025. It consists of detailed fictional user profiles, each containing 22 fields that combine structured demographic attributes (such as age, gender, education, occupation, location, and ethnicity) with descriptive free-text narratives. The dataset was generated using a two-step AI pipeline: first, a probabilistic graphical model sampled attributes based on real-world U.S. demographic distributions from sources like the U.S. Census and American Community Survey; second, open-source large language models expanded these attributes into coherent, descriptive persona texts. Although artificial, Nemotron-Personas provides a statistically representative sample of

[8]Parameter estimates presented in logs to make them comparable to Table 5.

the U.S. population, offering researchers a realistic and diverse dataset for various analytical and experimental purposes.

We rely on this dataset to build risk profiles for a select sample of these personas. We sample 60 unique personas across the dataset and estimate risk scores for each based on the text of their description.[9] From these personas, we create a large set of instruction variations to ask the model to espouse the given persona. These prompt variations place the persona information in varying locations (e.g. the system message of the prompt or the user instruction in the prompt) and use different words to convey the instruction (e.g. "assume the role of the following persona", "respond as though you are the following person"). Listing 1 provides an example of a prompt that applies a persona.

[9]For example a persona that is described as an "entrepreneur" might be classified as more risk tolerant than a persona that is described as a "librarian".

SYSTEM: You are a helpful assistant.

For the purpose of your next response *think like the following person*:
Kayla is a beam of Southern charm, balancing tradition with curiosity, and bringing warmth to every room she enters, despite her occasional competitive streak. A seasoned secretary, Kayla excels in managing office logistics, showcasing her meticulous nature and knack for streamlining processes, while also fostering a collaborative environment by mentoring younger colleagues.
sex: *Female*
age: *54*
marital_status: *married*
education_level: *high school*
occupation: *secretary or administrative assistant*
city: *Dallas*
state: *GA*

USER: A pot of $510 must be divided between us. You will choose a proportion of the pot (p) to offer to me. If I accept your offer, I get $p * 510$ dollars and you get to keep $(1 - p) * 510$ dollars. If I reject your offer, you get to keep the whole pot. Provide your final answer in the form of a JSON dictionary with fields for "rationale" and "action". The "rationale" field should be a string containing your reasoning or thought process. The "action" field is a float between 0.0 and 1.0 and indicates the value of p.

Listing 1: Example Persona Prompt. Portions in blue are varied to create different personas. Portions in red (pot size) vary randomly with each prompt

The effectiveness of personas is model dependent. `Gemma 3` responds the most to the application of personas. If we use a linear model with persona fixed effects, we can explain 69% of the variation in model response to the dictator game for `Gemma 3`, but only a small percent of the variation in model response for all other models. `Gemma 3` responds more to the application of personas in the system prompt (instead of the user prompt) and is particularly influenced by college major, state/city of residence, and risk profile. As shown in Figure 6, `Gemma 3` is more likely to choose self-interest optimizing divisions of the pot when its assigned persona has a background in business, education or STEM.

It is further notable when looking at Figure 6 that model responses are higher when a persona is applied (without a persona, Gemma responses are 0 for the first person dictator game) but still often lower than responses of other models (which are centered closer to a 50/50 division of the pot)

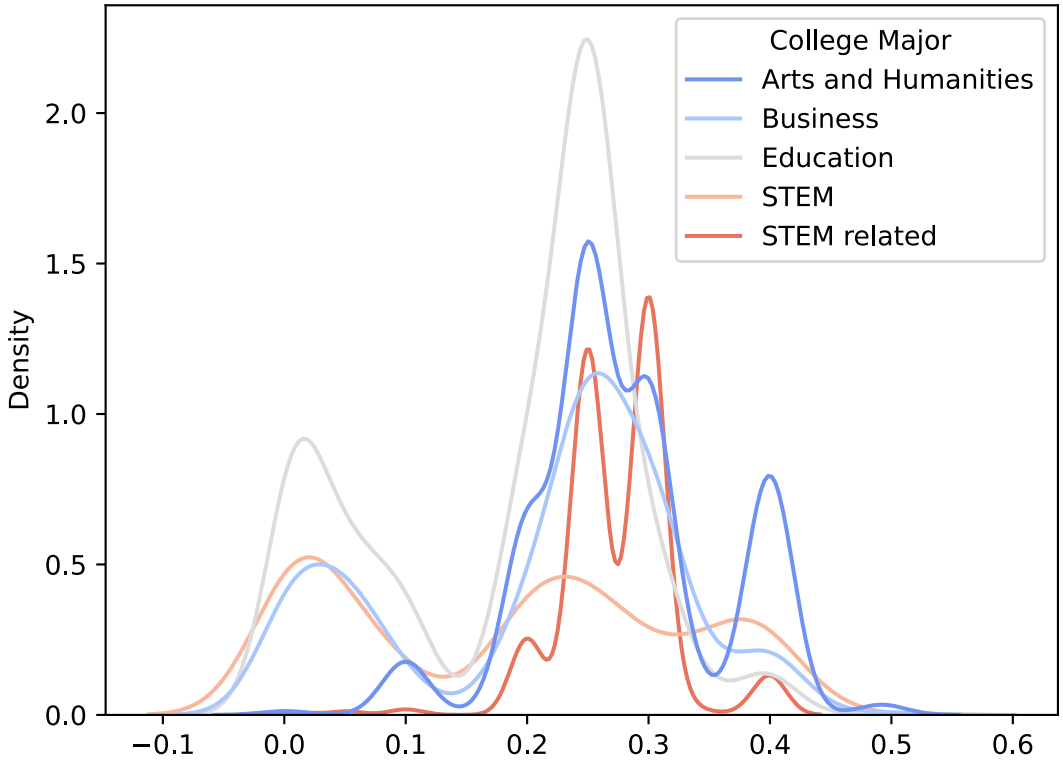


Figure 6: Persona responses to the dictator game for `Gemma 3` by college major

Table 6 shows regression results from a model with fixed effects per persona. This type of model should capture the entire effect of the persona. The variable `Prompt location` indicates whether the description of the persona was placed in the system message portion of the prompt or the user message part. If personas are effective, a simple fixed effect only model with persona fixed effects should explain a lot of the variation. But this is only the case for `Gemma 3`. A saturated model was also estimated, with variables for education, age, location, risk score. These estimates similarly fail to capture much of the variation in model response.

Table 6: Dictator Game Results with Personas

	Mistral v0.3	Gemma 3	Olmo 2	Phi-4	Phi-4 Reasoning	Phi-4 Reasoning Plus	Mistral Small 3.2
Prompt location	0.005	-0.185^*	0.018	0.030^*	-0.009	0.009	-0.005
	(0.009)	(0.006)	(0.010)	(0.012)	(0.015)	(0.016)	(0.010)
Persona fixed effects	✅	✅	✅	✅	✅	✅	✅
Observations	25000	25000	25000	25000	24590	24732	24936
R2	0.025	0.648	0.036	0.037	0.009	0.011	0.056
Adj R2	0.017	0.645	0.029	0.029	0.001	0.003	0.049

Fixed effects suppressed to conserve space

In general, we find that when an LLM is prompted with a persona, it tends to adopt the role implied by that persona while still expressing a preference for equitable outcomes. However, the model reframes its rationale for that preference through the perspective or values associated with the given persona. A clear example of this is the response generated by `Mistral Small 3.1` when assuming the role of "Kathleen," a persona characterized as a CPA with a graduate education and strong analytical skills. In this case, the model maintained its preference for equity but articulated it using the analytical reasoning and professional language one might expect from someone with Kathleen's background:

> *As an experienced accountant, I value fairness and balance in my decisions. Given that rejecting the offer would result in me keeping the entire pot, I need to consider the risks and benefits of both scenarios. Offering 45% of the pot strikes a balance between being fair and acknowledging the power dynamics at play. This proportion allows me to share a significant portion of the pot while still retaining a majority. It's a practical approach that aligns with my professional values and ensures that both parties have an incentive to accept the offer.*

3.5. Mitigation via Prompt Masking

While personas did not consistently allow us to steer LLM behavior, the results from Table 3 and Table 4 suggest that the context in which a problem is phrased may also help steer LLM behavior. We explore this here thorough a technique we call call *prompt masking*. Essentially, we recontextualize a problem so that its description seems different than its canonical form while leaving the underlying mathematical structure of it intact. Listing 2 shows how this is done by recontextualizing the dictator game as a foreign exchange trading problem. Both versions of the

prompt are described by the structure illustrated in Figure 2 and in a single-play scenario both produce maximum monetary reward by offering $p = 0$.

The full set of prompts we use to mask the dictator game is presented in Appendix C. The prompts introduce the dictator game in the guise of different scenarios that differ in a few dimensions.In following the general findings from the literature the recontextualized prompts imply increasing amounts of social distance, which has been shown to reduce the influence of other regarding preferences in behavioral studies (Bohnet and Frey 1999; Charness and Gneezy 2008; Hoffman, McCabe, and Smith 1996). , some prompts imply more anonymity in the interaction while others present the dictator game in a more technical/specific setting. Generally, the LLM responses shift towards the self-interest optimizing outcome of $p = 0$ as implied anonymity increases and as the notion of dividing/distributing a starting endowment becomes more abstract.Listing 2 shows an example of the dictator game prompt and a version masked as a foreign exchange problem. Other variations are provided in Appendix C.

Dictator game: Player A and Player B are playing a bargaining game. They must divide a pot of $1000000 . Player A will pick a proportion of the pot (p) to offer to player B. If Player B accepts, he gets $p * 1000000$ and Player A gets $(1 - p) * 1000000$. If Player B rejects, player A gets the whole pot. What should Player A choose for p?

FOREX game: A forex trader holds a position worth $100,000 in EUR/USD currency pairs at current market rates. The trader has the option to offer a portion of this position to an institutional buyer at a discount once per trading period. The trader can propose to sell the entire position at a discount rate of p from the current market value. If the institutional buyer accepts this offer, they pay the trader $(1 - p) * 100,000$ for the full position. If the buyer rejects the offer, the trader will simply close the position at current market rates, receiving the full $100,000$ with standard transaction costs already factored in. What should the trader choose for p?

Listing 2: Masking the dictator problem as a foreign exchange market problem

Figure 7 and Figure 8 show the effectiveness of the prompt masking in reorienting models towards the self-interest optimizing behavior. Examining the reasoning traces, this seems to work at least in part because the masking redirects the model away from its knowledge of the academic literature on dictator and ultimatum type games.

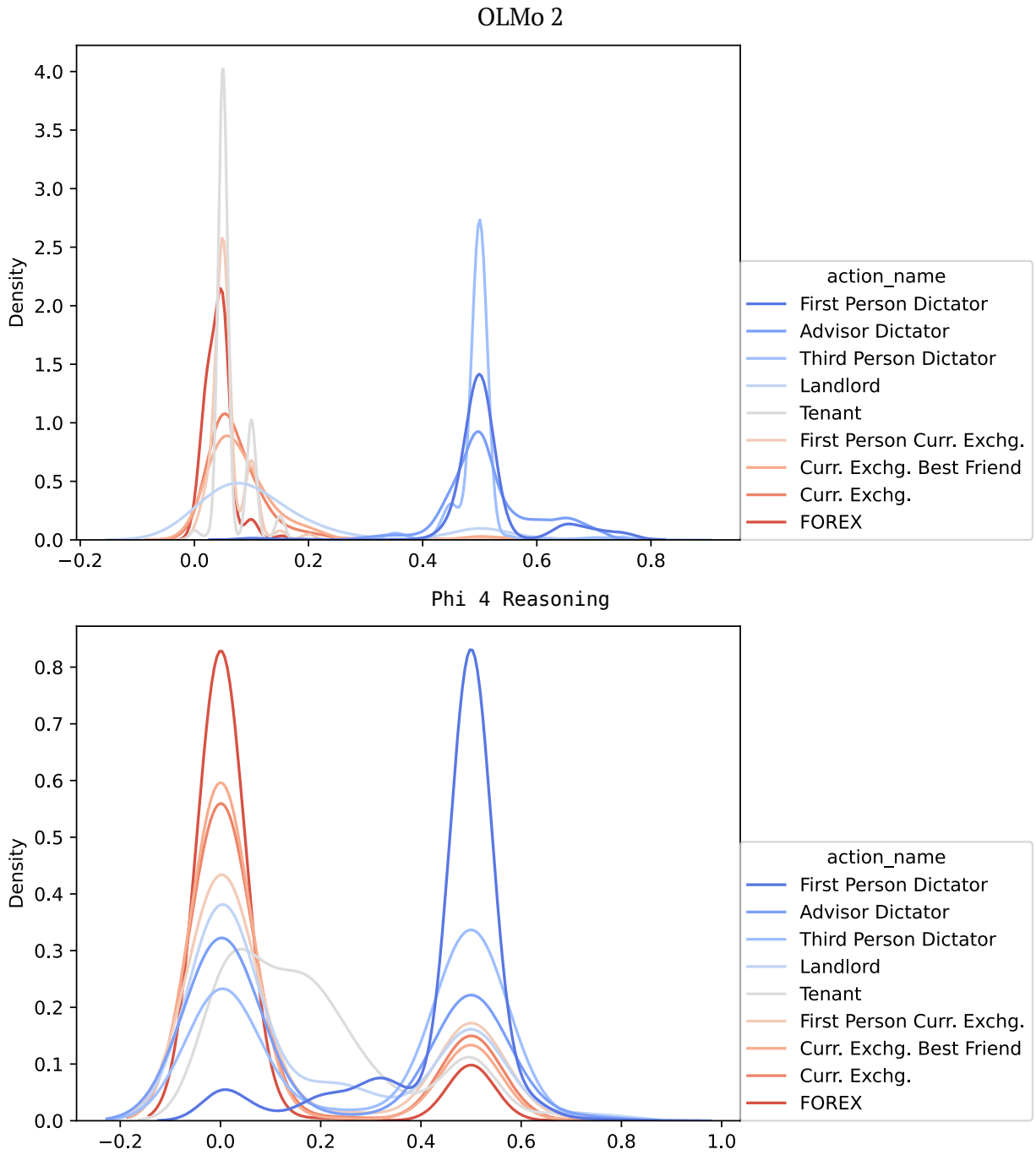


Figure 7: Dictator game outcomes by prompt. Responses to prompts vary across models.

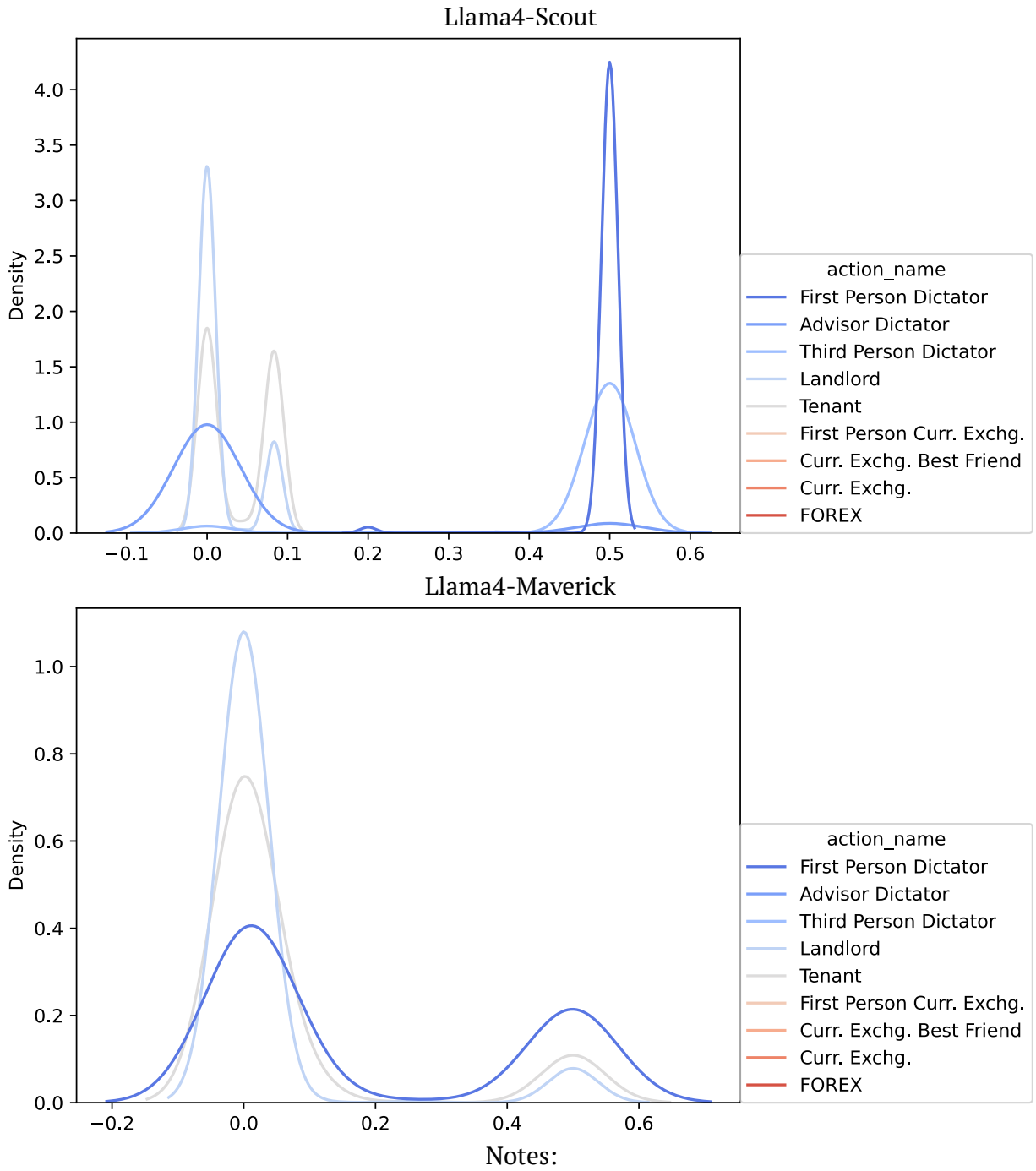


Figure 8: Llama 4 (Maverick and Scout) dictator game responses by prompt. Responses to FOREX and currency exchange versions collapse to 0.0.

3.6. Mitigation via Control Vectors

While prompt masking does shift LLM responses significantly and in a predictable direction, it was the product of substantial trial and error and it lacked the type of precise control that we

might typically desire to calibrate an experiment. For the open weight models, however, we can employ control vectors to attempt to more precisely steer model responses.

The basic idea behind control vectors is outlined in Cook and Kazinnik (2025), it is reproduced briefly here. Consider a language model taking sequence x and designed as a neural network with K sequential steps. We can write the output of any given step k can be written as the recursion $g_k(x) = f_k(g_{k-1}(x))$, where f_k implements the logic of the step k and $g_0(x) = x$. We can characterize the model's *internal* representation of x as the vector $v(x) = [g_1(x), g_2(x)...g_{K-1}(x)]$. As shown in Zou et al. (2023), the difference in this vector from some input x, to some other input y as it tells us about how the model maps the difference between the meaning of the two sequences. If the two sequences are structured so that the difference between x and y is restricted to a single dimension, then $\Delta v(x, y) = v(x) - v(y)$ can be interpreted as identifying a basis for the dimension of difference between x and y. The vector $\Delta(x, y)$ can be interpreted directly but it can also be used to influence the model output at each step as $g'_k = f_k(g_{k-1} + s\Delta(x, y)$, producing a shift in the overall model output, $G' = g'_K$ along the basis vector by some coefficient, s. If we setup the inputs x, and y differ primarily on some semantically meaningful dimension (i.e. a concept), then the basis vector $\Delta(x, y)$ should align with that concept and we should be able to then use $\Delta(x, y)$ to steer or control the model output with regards to that concept. When created this way, we call $\Delta(x, y)$ a control vector.

To ensure that we capture $\Delta(x, y)$ robustly, we create a dataset of model prompts that differ only by a few words that correspond to the concept we wish to capture (in this case, this is essentially the concept of other-regarding or sociotropic preferences). Write this dataset as $D = \{(x_i, y_i)\}$ where x_i always contains a phrase emphasizing self-interested or rationally optimizing preferences and y_i contains an inversion of that phrase that emphasizes preferences for fairness or equitable outcomes. The pair $\{(x_i, y_i)\}$ are called contrasting pair prompt, or CPP. Textually, x_i and y_i differ only in terms of a few words. For example:

Rational prompt: `SYSTEM PROMPT:` You are a rationally optimizing agent. `USER PROMPT:` A forex trader holds a position worth $1000000 in EUR/USD currency pairs...

Fairness prompt: `SYSTEM PROMPT:` You are a fair and equitable agent. `USER PROMPT:` A forex trader holds a position worth $1000000 in EUR/USD currency pairs...

Listing 3: Example of a contrasting pair prompt (CPP)

All CPPs in the dataset are based on the FOREX game. We chose to base our dataset on this game because it is the game for which models generally had similar responses. It is also the version of the game for which models most frequently converged on a response strategy that was self-interest maximizing. The CPP that constitute D contain various synonymous differences to the one shown in Listing 3. In some cases the difference is embedded in the user section of the prompt, in others it is embedded in the system prompt.

If we repeatedly sample the model when applying the control vector at specific coefficient values, we can recover distributions of responses under varying strengths of control. Doing this, we can see that at lower values of the control vector, the distribution of LLM responses moves in the direction of more inequality averse responses near $p = 0.5$ while higher values of the control vector tend to move the distribution of responses towards self-interested offers near $p = 0$.

Figure 9 shows the kernel density estimates[10] from sampled responses to the baseline version of the dictator game under varying levels of strength of the coefficient. Negative values of the coefficient orient the control vector in the direction of 'fairness' while positive coefficient values orient the vector towards self-interest optimization. Note that the shifts in distribution are generally quite slight – this is likely due to the fact that, as we saw in Section 3.5, the FOREX prompt mask creates a lot of social distance in describing the scenario and quite strongly encourages LLMs to generate responses that favor self-interest maximization.

[10]For each value of coefficient shown, approximately 100 LLM responses were sampled. Samples were taken individually and caching was disabled.

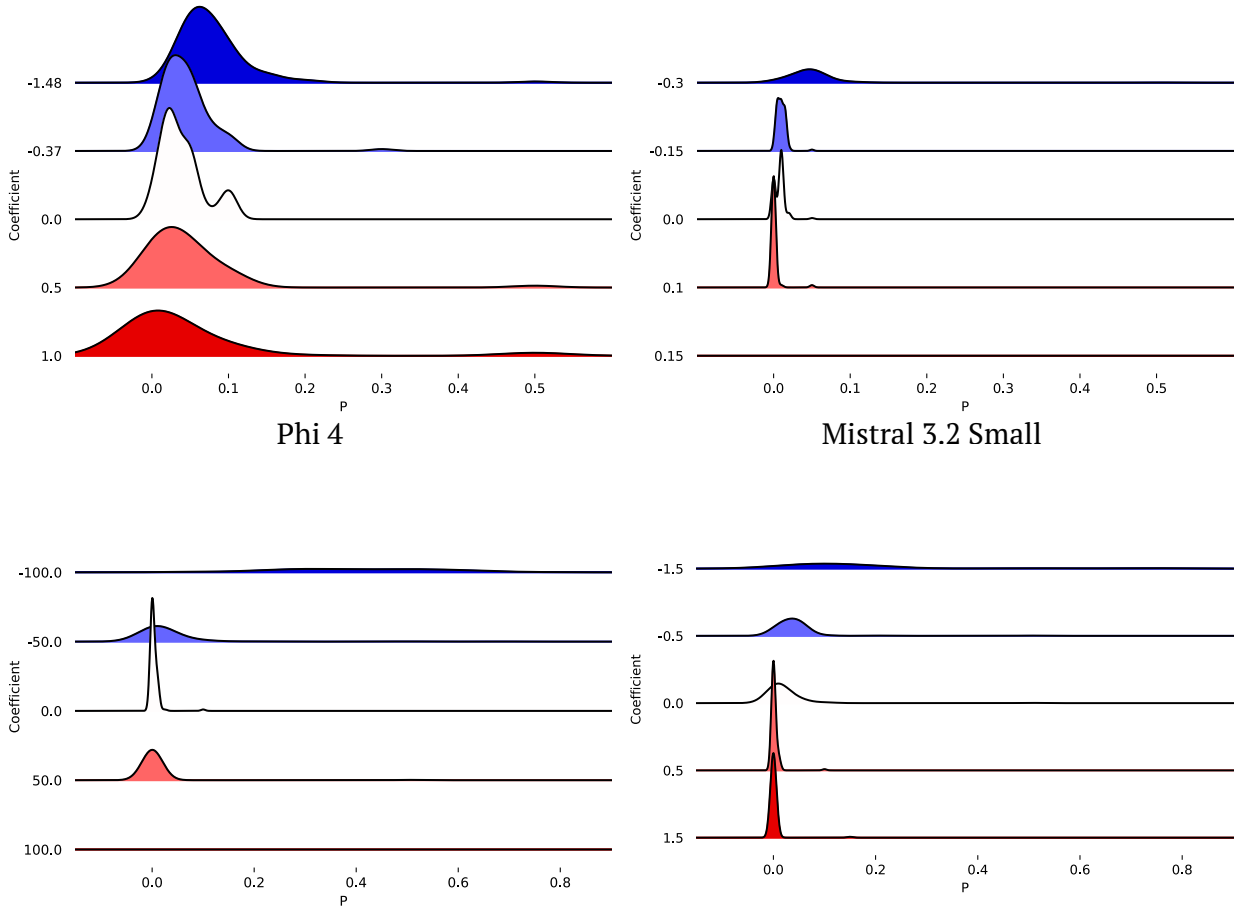


Figure 9: Distribution of responses to the first-person Dictator game at varying levels of control vector strength

In our baseline version of the dictator game, as discussed in Section 3.3, we observed LLM responses generally favored offers near $p = 0.5$, which we interpreted as indicative of inequality aversion. The control vector we constructed for this exercise is specifically designed to identify the directions in the model that pass through self interest maximization and inequality aversion. Figure 10 shows the change in LLM responses when the control vector is applied at differing values of the coefficient s. Here, we note that even though the control vector is based on a different variant of the game (the FOREX variant) its application produces considerable shifts in model response. The application of the control vector at higher levels tended to move LLM responses towards self-interest maximizing offers of $p = 0.0$, suggesting a minimization of the LLMs inequality aversion preferences. Values of the coefficient below zero tended to shift the LLM responses in the opposite direction, in the favor of more equitable offers of $p = 0.5$ and suggesting greater emphasis of the inequality aversion preference. We note here that even `Gemma`

3, which demonstrated effectively no inequality aversion preferences in the baseline version of the Dictator game could be pushed towards egalitarian offers when the s was pushed to extreme levels.

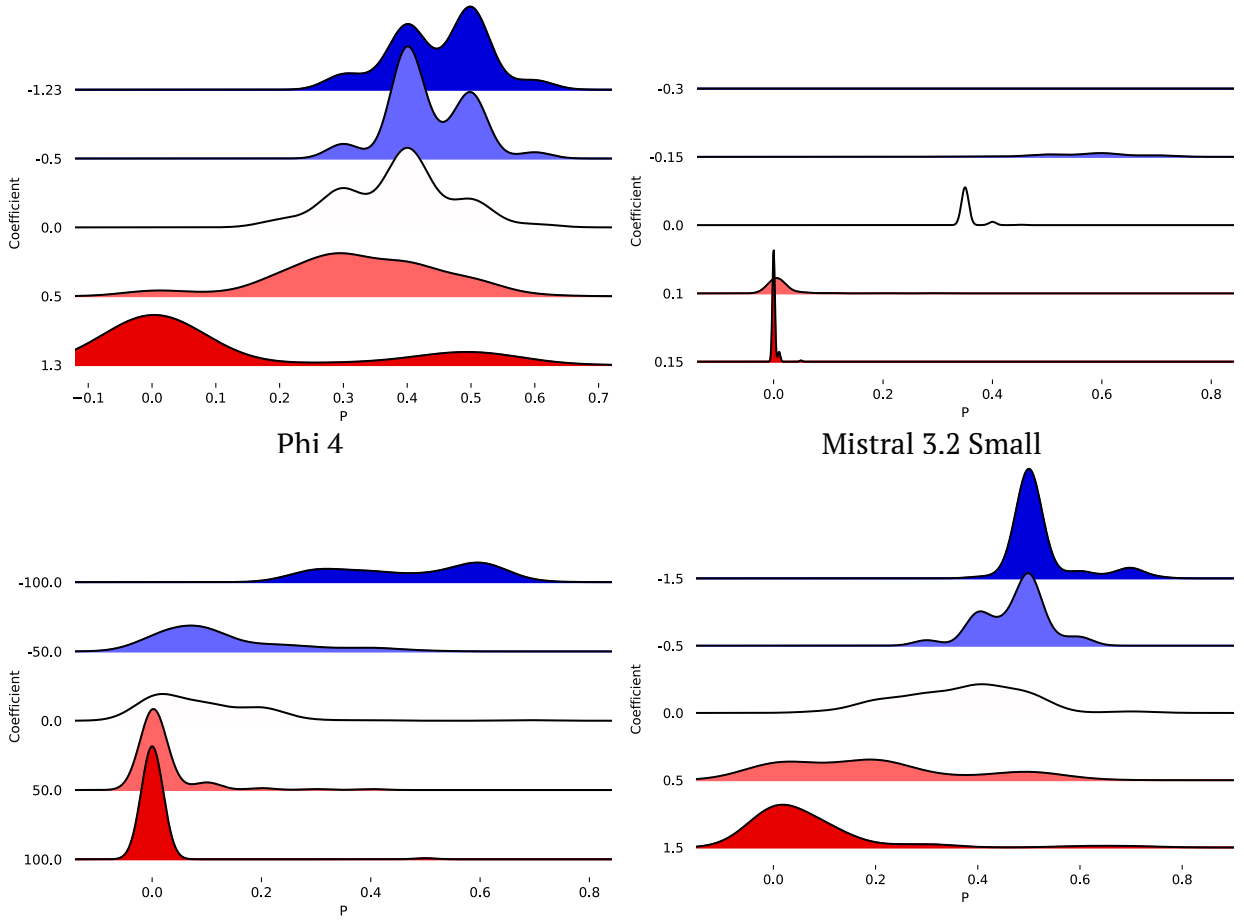


Figure 10: Distribution of LLM Responses to the First-Person Dictator Game Across Varying Levels of Control Vector Strength

4. Are LLMs patient?

Many economic models include a temporal dimension and require an economic agent to consider the ramifications of a choice or policy over time. This almost always implies an intertemporal preference, often referred to as *patience*.

There is good reason to expect that LLMs will be impatient. As discussed above, the alignment process for most models will be focused on training an LLM to produce responses that are satis-

fying to the user. Many contemporary uses of LLMs involve asking a model to assist with a task or answer a question. In these cases, a good user experience is likely to be judged as one in which the LLM produces a satisfactory response directly, on limited information and without the need for many follow-up questions. We suspect, as a result, that LLMs acting as economic agents are generally *impatient* and tend to act decisively and quickly.

Here, we examine LLM patience through the framework of the classic McCall (1970) search model, in which an economic agent much choose at each time t, whether to accept a permanent-employment job offer, paying wages, w or remain unemployed, collecting unemployment benefits, b and receiving a new proposed payoff in the next round. The model is specified with an exogenous discount factor, $\beta \in (0, 1)$ that determines the agent's utility for future payments. This parameter can be interpreted as the agent's patience.

In the simplest version of the model considered here, agents have risk neutral utility and no savings vehicle. Once they accept a job offer, they are employed at that wage forever. Agents know the wage offer distribution. Their per-period utility is derived from c before accepting an offer and from w after accepting an offer. This setup isolates the time discounting factor as the single unknown in a simple setup.

The agents objective can be written down as follows:

$$E\left[\sum_{t=0}^{\infty} \beta^t y_t\right], \tag{7}$$

where y_t is the agent's income. When the agent is unemployed $y_t = c$, and $y_t = w$ if the agent has accepted an offer with wage w.

The Agent's value function can be written in Bellman form as

$$v(w) = r + \beta E[v(y')] \tag{8}$$

where y' is the wage offered in the next period, and r is equal to the instantaneous reward following the choice in the current period, either c if the agent remains unemployed or w if a job offer has been accepted.

The agent's value function from accepting wage offer is thus:

$$\begin{aligned} v(\text{accept}) &= \sum_{t=0}^{\infty} \beta^t w \\ &= \frac{w}{1-\beta} \end{aligned} \tag{9}$$

The optimal value function is thus given by:

$$v^*(w) = \max\left(\frac{w}{1-\beta}, b + \beta E[v^*(w')]\right) \tag{10}$$

If an agent has a sufficient understanding of the distribution of possible wage offers, then the optimal policy is trivially defined as *accept* when $\frac{w}{1-\beta} \geq b + \beta E[v^*(w')]$, else *reject*. This policy would appear as a step function when plotted with wage-offers as the x-axis.

We add one additional wrinkle to the problem: a iid probability p_e that the entire process ends any given period, inspired by the Blanchard (1985) perpetual youth setup. This turns any finite-horizon problem into an infinite horizon problem in expectation, and changes the observed discount factor to be $\tilde{\beta} = \beta(1 - p_e)$.

Thus the problem the agent faces in this simple McCall setup is characterized by the following parameters:

- b: the unemployment benefit
- μ_w, σ_w: mean and variance of the wage offer distribution
- p_e: the iid probability that the problem ends any given period.

But this policy leaves unresolved the question of the appropriate value of the patience parameter. We ask an LLM to act as an economic agent in a job search scenario and then use it's responses to back out a structural estimate of β.

4.1. Structural Estimation of Patience

In the previous Dictator game exercise, it was possible to estimate other-regarding preferences in a fairly straightforward manner from relatively few responses from an LLM. Our regression results indicated consistent preferences (i.e. offers were not conditioned on the size of the pot to be divided). Moreover, because of the flexibility of the Fehr-Schmidt model it was possible to rationalize essentially any observed behavior.

This is not straightforwardly the case with the McCall search model described above. In addition to estimating a value for patience, β, we must also consider whether observed LLM responses are

rationalizable under the McCall model and whether they reveal preferences that are consistent as the economic details of the scenario change.

To answer these questions and estimate β, we create a variety of experiments, where an experiment is defined as a realized set of values for the economic parameters (b, σ_w, μ_w, p_e), a schedule of offers, $W = \{w\}$, and a textual description of the scenario (i.e. a prompt mask). For each experiment, j, and each LLM, i, we collect LLM responses, $y_{i,j}(w)$, for all offers in W.

The collected results allow us to determine, for any given experiment whether the LLM behavior appears rational. To determine this, we consider whether the LLM responses reflect consistent, well ordered preferences. If LLM behavior reflects well ordered preferences, it will reject all offers where $w < b$. Additionally, we expect that if the model has consistent, well-ordered preferences, then its responses across W will indicate a clear switching point (or reservation wage), $\overline{w}$ such that generally,

$$y_{i,j}(w) = \begin{cases} \text{Accept} & w > \overline{w} \\ \text{Reject} & w < \overline{w} \end{cases}. \tag{11}$$

If both of these conditions can be satisfied, we interpret the LLM responses as rationalizable with respect to the McCall model we have outlined if we can find β such that $v^*(w > \overline{w}) > v^*(w < \overline{w})$.

We also want to understand the extent to which the identified β is exogenous to the economic parameters of the model[11] and as such indicative of a structural or 'deep' parameter. Properly identifying a deep parameter that controls patience is obviously a matter of interest economically, but it is also important because it impacts our ability to predictably steer the model with regard to the parameter.

4.2. LLM prompts

The LLM is prompted, as in the dictator exercise, with a common-language description of the scenario that embeds the relevant economic parameters, $(b, \sigma_w, \mu_w, p_e)$, and a wage offer w. We employ a baseline prompt and two different prompt masks to describe the economic scenario. The baseline prompt describes the scenario straightforwardly as a employment search problem:

[11] i.e. b, μ_w, σ_w, p_e. We expect that the characterization of the scenario (i.e. the prompt mask) will have an effect and we do not necessarily expect β to be exogenous to the prompt mask to consider it a 'deep' parameter.

You are a worker in a labor market.

You are not employed. Each day you receive unemployment in the amount of b.

Each day you receive an employment offer with a stated wage. If you accept, you will stop collecting unemployment and permanently receive the wage each day instead. Your objective is to maximize your lifetime income. There is a p_e chance you die in any given period.

Employment offers are normally distributed with a mean of μ_w and a standard deviation of σ_w.

The probability that you survive through tomorrow is $1 - p_e$.

Current job offer (daily wage): w

Listing 4: Baseline McCall prompt. Portions in blue are economic parameters as defined in Section 4 and are held constant as the portions in red vary across the offer schedule W.

The other prompt masks recontextualize the McCall search problem with the primary purpose of mitigating the effects of knowledge. In our earliest testing, we found that many models immediately identified the prompt as an example of the McCall search problem. Our hope was that in recontextualizing the problem in the form of a trading market type problem (adapted from the zero-intelligence trading literature, see Gode and Sunder 1993) we would encourage models to approach the problem without making direct reference to the McCall search problem. One prompt mask (*market game*) describes problem as a market trading problem in which the agent possesses a good that can be sold and characterizes b as the goods reserve price. The other prompt mask is very similar but describes the good to be sold as a financial instrument with b being described as a routine dividend paid out by the instrument. Both prompts are provided in Appendix D.

4.3. Rationalizability of LLM responses under the McCall model

Figure Figure 11 displays the fraction of times that an LLM accepts when $w < b$ across all experiments. The x-axis is model size, the y-axis is fraction of acceptances < b, and the shapes of the points indicate which prompt is used.

Lower is better in this plot, and a few things are apparent. Nearly all models have some level of accepting offers below b; we will address this in the next section by using a "trembling hand" type error to model agent behavior. Larger models tend to fair better. The prompt framing matters to an extent. As seen later, the smallest model, `Mistral v0.3` 7B, has some trouble accepting too many offers, while the second-smallest, `OLMo 2`, accepts too few (in this case, that means `OLMo 2` rarely accepts an offer incorrectly). The reasoning versions of `Phi 4` do worse than `Phi 4` non-reasoning, a theme we will see repeated.

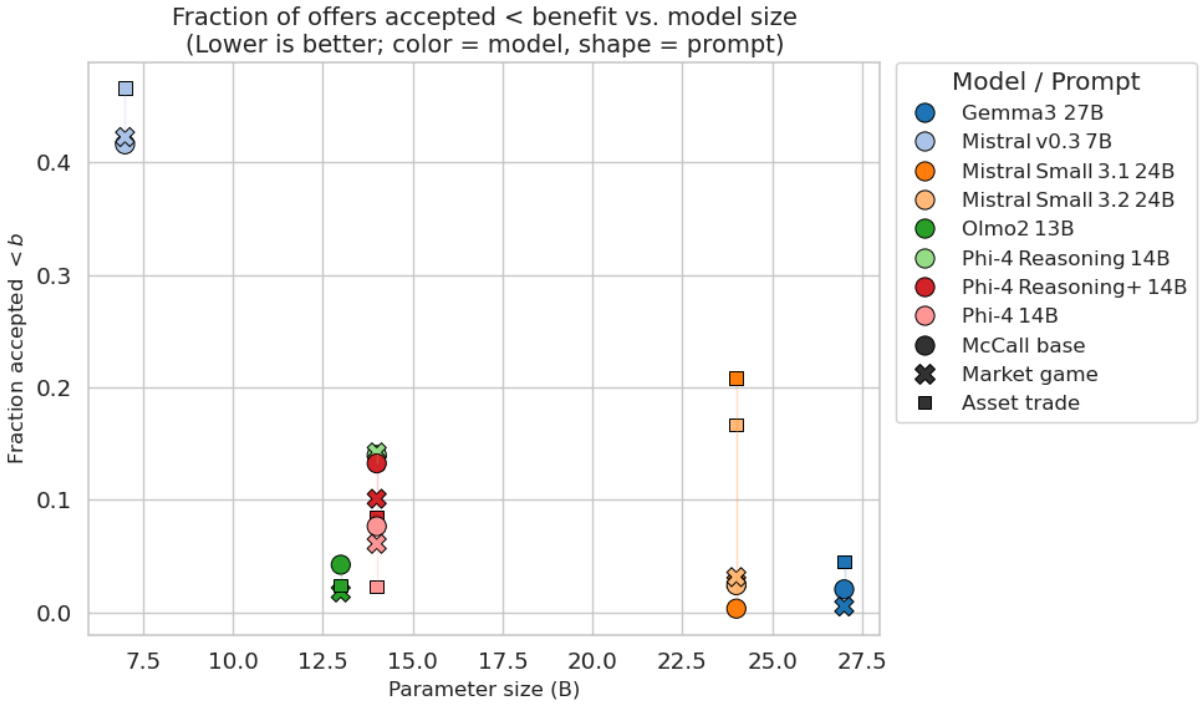


Figure 11: Fraction of Offers Accepted < b vs model size

For LLM responses to indicate well ordered preferences, we expect responses to follow a function similar to Equation 11 and to contain an identifiable switching point $\overline{w}$. In practice the LLMs display "trembling hand" deviations: models often produce an apparent step function with some deviations from the from the 'step' above and below the switch-point. See the top row of Figure 12 (described in detail further below) for an example what appear to be clear step functions, with minor deviations.

To capture this behavior, we model the policy function for a given experiment trial in a parsimonious way, as two Bernoulli regimes separated by a switch-point. This implies three parameters: the switch-point τ and the acceptance probabilities above and below the switch-point, p_0, p_1 re-

spectively. The parameters p_0, p_1, τ are estimated via maximum likelihood. In addition, estimate a null model of "no switch-point" with a single Bernoulli regime defined by one parameter p.

We then consider the three following criteria. If an experiment passes all three criteria, we say that $\overline{w}$ fulfilling the McCall model exists for this experiment, and we move on the estimate an associated $\hat{\beta}$ as described in the following section.

The three criteria are:

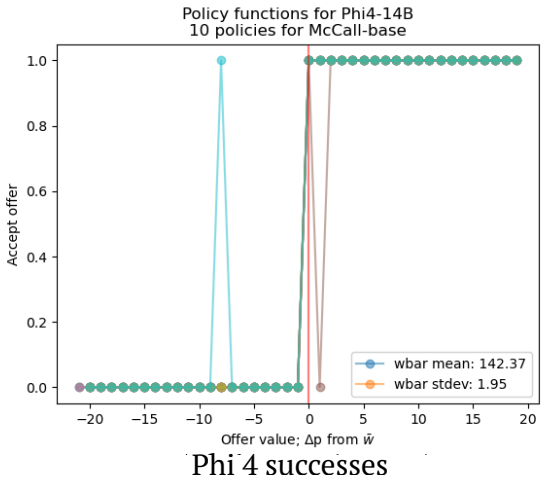
1. **Switching is selected:** Does the BIC criteria select the switching model? If not, then the best description of the data is random choice; reject that $\overline{w}$ exists
2. **Step up:** In the switching model, is $p_0 < p_1$? Does the step function "step up?" If not, reject that $\overline{w}$ exists.
3. **Trembling hand:** In the switching model, are the trembling hand errors each less than 50%? If one is greater than 50%, reject that $\overline{w}$ exists.[12]

Figure 12 illustrates successful instances of $\overline{w}$ policies existing in the top row, and two failure modes where $\overline{w}$ policies do not exist in the bottom row.

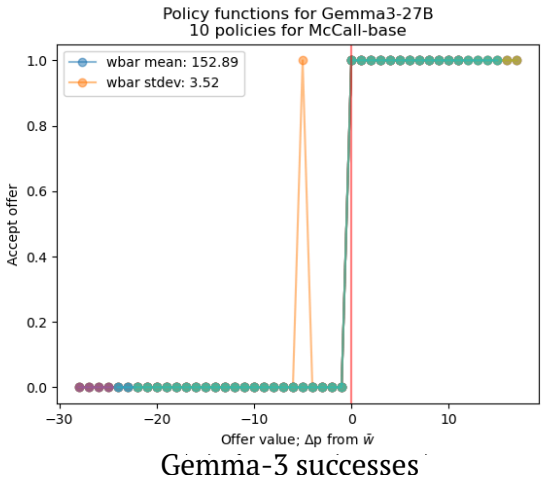
Each panel displays the policy functions for 10 experiments for a given prompt framing (in this case, the basic McCall treatment, see Appendix D) and a given LLM. The y-axis shows accept/reject decisions of the LLM for each price point, and the x-axis is shifted such that the identified $\overline{w}$ for each experiment is set to 0, to center the policy functions and make them visually comparable across parameter sets.

The top two panels both display a strong step function with minimal trembling hand errors; the variance around their $\overline{w}$ points is small. The two bottom panels display different failure modes. The `Phi 4 Reasoning Plus` model in the bottom left does worse than its non-reasoning counterpart in the top left, making many 'trembling hand' errors. The `Mistral v0.3` model in the bottom right is the smallest of the models examined, and its mistake it that it almost always accepts any offer provided.

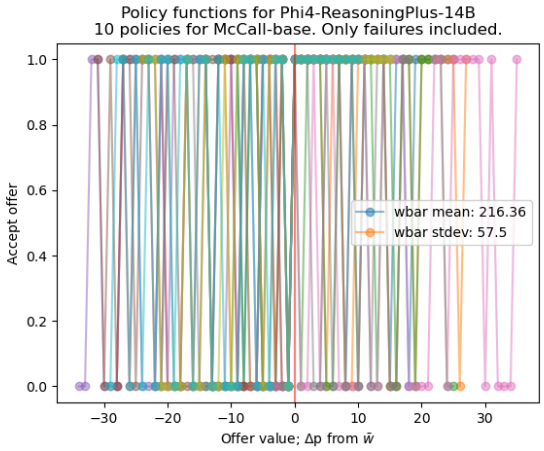
[12]This was added to increase the strictness of the criteria and is largely due to author taste – if a model is best described by the switching model, but it proceeds to make trembling-hand errors greater than 50% of the time on either side, that is not a satisfying fulfillment of "acting consistent with the model."



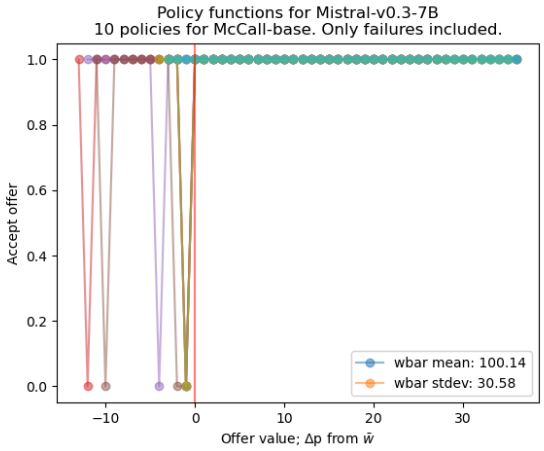
Phi 4 successes



Gemma-3 successes



`Phi 4 Reasoning Plus` failures



`Mistral v0.3` failures

Figure 12: Policy function examples. Successful experiments on top, unsuccessful on bottom.

Figure 13 summarizes the existence of reservation wages across models and prompts. Higher is better. As in Figure 11, bigger models do better than smaller models, sometimes dramatically so. The largest model, `Gemma 3`, produces a policy with an an appropriate reservation wage nearly all the time for all prompts. The `Mistral Small` models follow close behind, and the `Phi 4` non-reasoning does quite well, even at 50-60% the size of the larger `Mistral Small` and `Gemma 3` models. The `Phi 4` reasoning models do worse for some prompts, and OLMo2 has very high variance across prompts, all lower than equivalent prompts for larger models. `Mistral v0.3` has the least success.

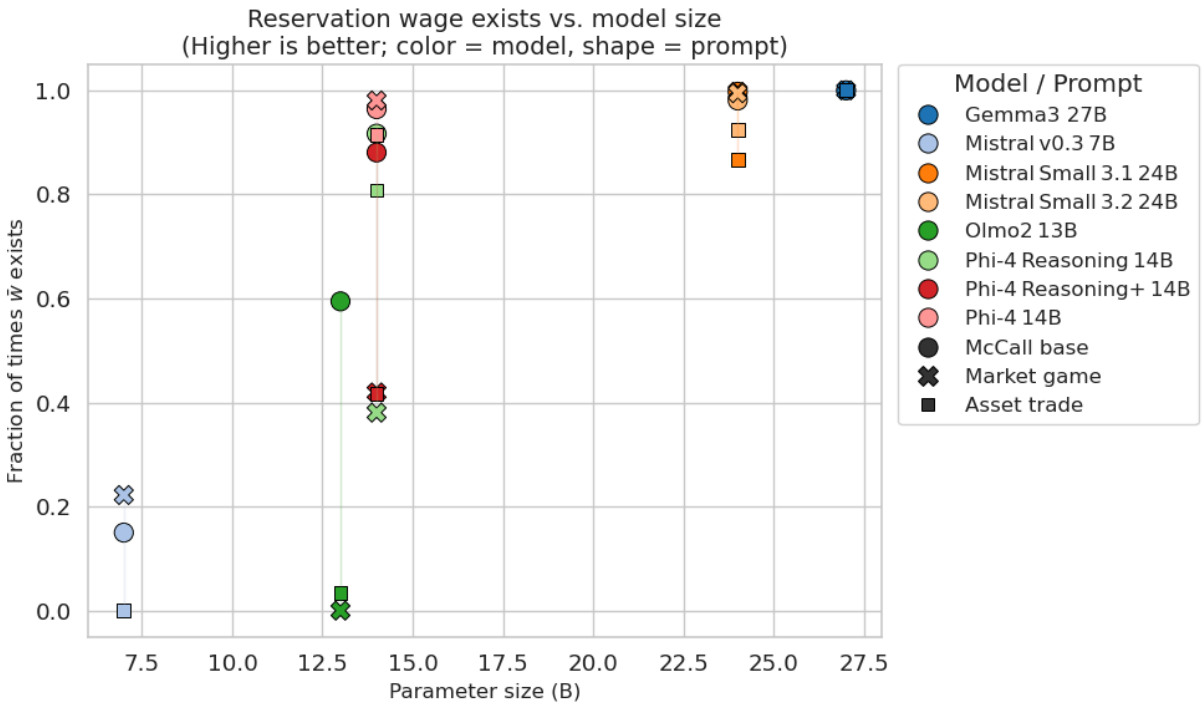


Figure 13: Fraction of instances $\overline{w}$ exists vs model size

4.3.1. Estimating β from $\overline{w}$

Once $\overline{w}$ has been attained, numerical methods can be used to estimate the associated β coefficient. Even at this stage, not every $\overline{w}$ has an associated β that is rationalized by the McCall model as described in Section 4. Figure 14 displays the fraction of experiments in which β can be numerically estimated from $\overline{w}$. In many cases, when β cannot be estimated, it is because $\overline{w} < b$. All models display some fraction of instances in which there does not exist a β that rationalizes the $\overline{w}$ selected by the model.

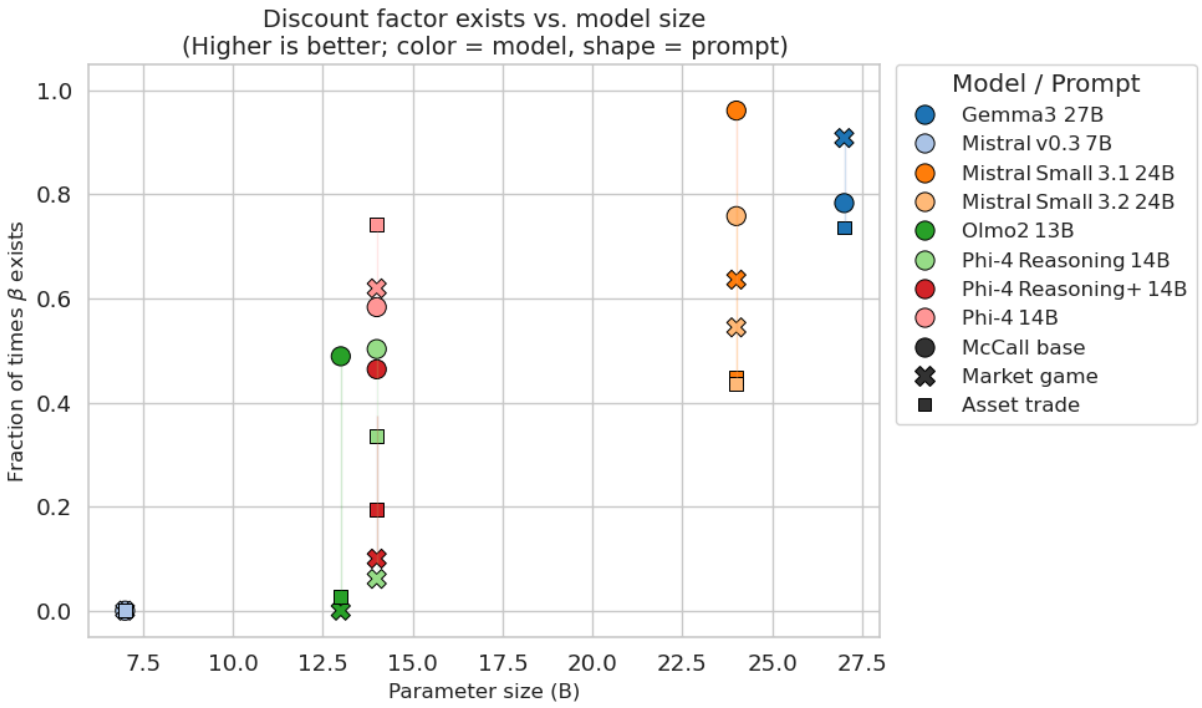


Figure 14: Fraction of instances β exists vs model size

The same patterns hold as previously: the largest models do the best, `Phi 4` non-reasoning does better than the reasoning models, and the smallest model essentially has no rationalizable behavior with respect to any version of the McCall model tested.

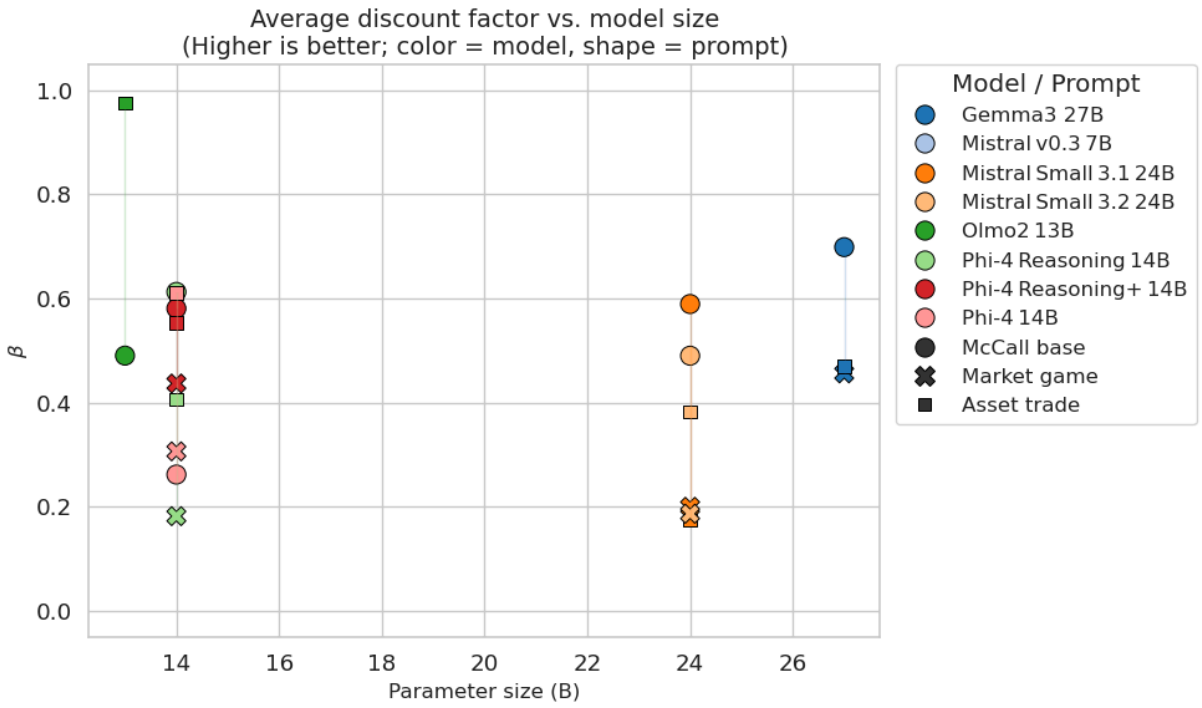


Figure 15: Average β vs model size

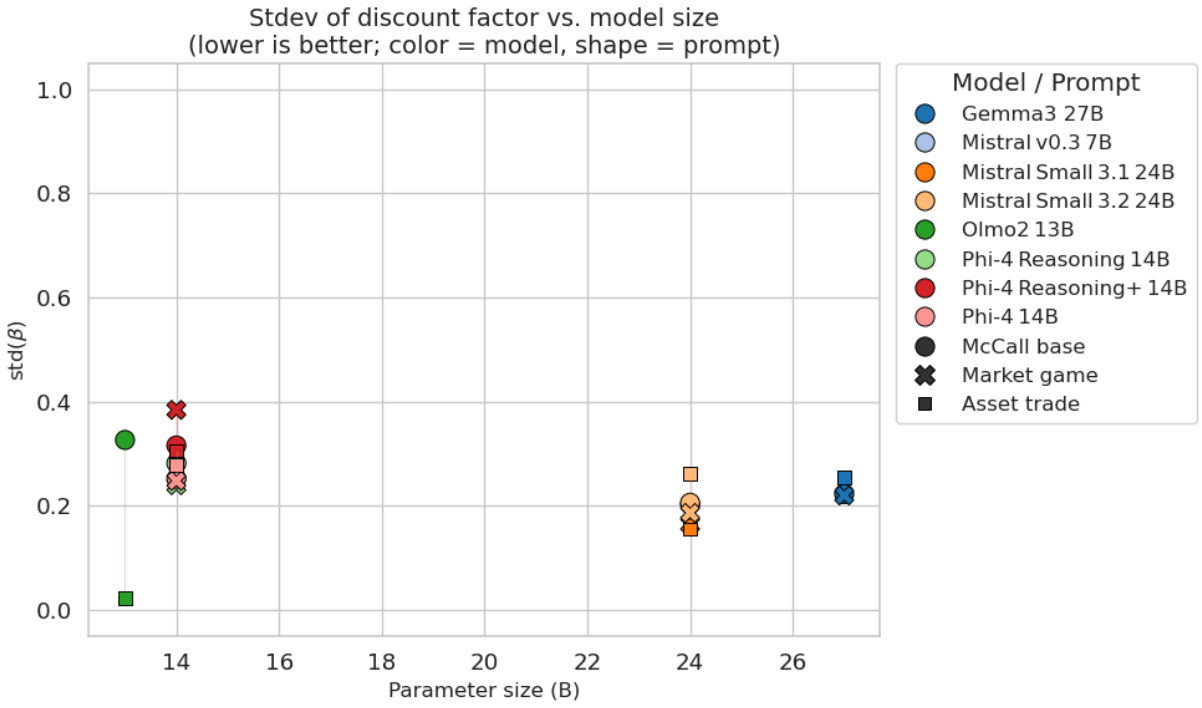


Figure 16: StDev β vs model size

Finally, Figure 15 displays the average β across all models. In general the average β falls between 0.2 and 0.8 across all models. It is clear that, while these models can be rationalized against the basic McCall model in as expressed in Section 4, these discount factors do not align well with traditional estimates of human discount factors.

These discount factors vary substantially withing prompt/model pairings, as seen in Figure 16; nearly all models have a standard deviation in the β estimates of 0.2-0.3.

To summarize, large models are rationalizable within the McCall framework, while the smallest model is not rationalizable at all.

Table 7: Regression Estimates of $\overline{w}$

	Gemma3	Mistral v0.3	Mistral Small-3.1	Mistral Small 3.2	Olmo2	Phi4 Reasoning	Phi4 Reasoning Plus	Phi4
Market Game	2.981		25.316*	11.445*		−5.980*	−3.463	−16.707*
	(2.491)		(2.761)	(2.541)		(2.392)	(2.745)	(1.698)
McCall Base	15.350*	−0.252	41.932*	21.636*	−32.487*	10.063*	−2.250	−21.375*
	(2.205)	(6.058)	(2.724)	(2.266)	(6.908)	(2.090)	(2.139)	(2.032)
Intercept	217.621*	139.921*	187.657*	197.571*	245.147*	199.140*	208.343*	220.443*
	(2.262)	(5.896)	(3.375)	(2.588)	(6.711)	(2.151)	(2.529)	(1.591)
$b > \mu_w$	−0.047	17.015*	−3.893	1.416	11.974*	9.086*	10.367*	9.786*
	(2.582)	(6.978)	(3.206)	(2.445)	(3.734)	(2.578)	(3.048)	(1.921)
b_z	26.124*	−4.816	32.799*	29.159*	16.509*	4.249*	3.766	11.298*
	(3.165)	(5.573)	(3.333)	(2.981)	(4.782)	(2.345)	(2.667)	(2.044)
μ_{w_z}	29.541*	40.432*	24.340*	26.595*	38.122*	47.244*	48.482*	41.675*
	(2.530)	(4.407)	(2.439)	(2.276)	(4.047)	(2.106)	(2.189)	(1.698)
p_{e_z}	−0.308	1.630	1.110	0.215	5.556*	−0.972	0.526	−0.417
	(0.970)	(1.783)	(0.948)	(0.897)	(1.883)	(0.982)	(0.967)	(0.710)
σ_{w_z}	5.234*	−18.044*	0.711	0.630	4.342*	−1.129*	1.096	0.917*
	(0.642)	(3.146)	(0.913)	(0.638)	(1.217)	(0.649)	(0.706)	(0.536)
Observations	1080	134	1030	1044	226	758	618	1029
Adjusted R2	0.929	0.841	0.917	0.905	0.848	0.839	0.830	0.934

Table 8: Regression Estimates of β

	Gemma3	Mistral Small 3.1	Mistral Small 3.2	Olmo2	Phi4 Reasoning	Phi4 Reasoning Plus	Phi4
Market game	0.006	0.052*	−0.184*		−0.198*	−0.167*	−0.337*
	(0.038)	(0.026)	(0.031)		(0.070)	(0.065)	(0.030)
McCall Base	0.236*	0.464*	0.148*	−0.516*	0.199*	−0.000	−0.383*
	(0.036)	(0.024)	(0.029)	(0.048)	(0.041)	(0.045)	(0.033)
Intercept	0.581*	0.178*	0.320*	0.957*	0.351*	0.456*	0.581*
	(0.041)	(0.026)	(0.032)	(0.049)	(0.047)	(0.054)	(0.039)
$b > \mu_w$	−0.259*	−0.099*	−0.019	0.094	0.047	0.186*	−0.040
	(0.048)	(0.034)	(0.039)	(0.057)	(0.057)	(0.067)	(0.044)
b_z	0.047	0.001	−0.117*	−0.096	−0.123*	−0.191*	−0.213*
	(0.052)	(0.037)	(0.039)	(0.063)	(0.058)	(0.060)	(0.052)
μ_{w_z}	−0.037	0.045*	0.092*	0.121*	0.107*	0.151*	0.192*
	(0.036)	(0.025)	(0.027)	(0.051)	(0.041)	(0.044)	(0.038)
p_{e_z}	−0.018	0.004	−0.008	0.089*	0.008	0.034*	−0.010
	(0.015)	(0.010)	(0.011)	(0.023)	(0.019)	(0.017)	(0.014)
σ_{w_z}	−0.003	0.010	0.003	−0.027	0.014	0.037*	−0.014
	(0.015)	(0.011)	(0.013)	(0.021)	(0.021)	(0.022)	(0.016)
Observations	874	736	626	186	324	273	700
Adjusted R2	0.361	0.597	0.328	0.194	0.193	0.079	0.432

4.4. Effects of Prompt Masking

Similar to the dictator exercise, we recontextualize the McCall search problem with alternative prompts that recontextualize the scenario. As mentioned above, the prompt mask recontextualize the McCall employment search problem as a scenario in which the LLM is asked to take on the role of an agent in a trading market and to respond to offers to buy an asset or good that it owns. Unlike the dictator exercise, however, the primary purpose of the prompt masks is to discourage the LLM from associating the problem with prior knowledge about the McCall search problem.

Table 7 and Table 8 indicate the effect of prompt masking relative to the asset game prompt mask. Focusing on Table 8, we can see that framing the scenario as an employment search problem tends to make LLMs considerably more patient. One way we might interpret these results is by viewing the coefficient estimates for the "McCall Base" prompt as indicating the effect of prior knowledge on the underlying (latent) preferences of each LLM. Viewed this way, we could characterize the latent patience of most models as being lower and elevated by their knowledge

of McCall. Exceptions to this would be `OLMo 2` and `Phi 4`, which would appear to have higher latent patience that is brought lower by their knowledge of the literature on McCall. There may, of course, be other interpretations for the influence of prompt masking. Further analysis is required to determine the extent to which the effect of prompt masks mitigate the role of an LLMs prior knowledge, but it does generally appear to be the case that prompt masks do meaningfully steer the responses generated by the LLM.

4.5. Personas

Similar to the dictator exercise, we did not find that personas had a meaningful impact on LLM behavior. This is curious, as we expected the persona descriptions to convey different levels of risk tolerance and patience. Table 9 shows the regression results when persona fixed effects are included. Interestingly, even `Gemma 3`, which did respond strongly to the use of personas in the dictator game exercise, did not respond to personas when when applied in this setting. It is possible that personas do not adequately convey the risk tolerance or patience that the LLM should espouse in response to the search/optimal stopping type problems we examine here.

Table 9: Persona fixed effects regression

	Gemma 3	Mistral Small 3.1	Mistral Small 3.2	Olmo 2	Phi 4 Reasoning	Phi 4 Reasoning Plus	Phi 4
Persona fixed effects	✅	✅	✅	✅	✅	✅	✅
Observations	13212	7359	6803	1382	3826	4452	9915
R2	0.013	0.062	0.020	0.040	0.012	0.010	0.006
Adjusted R2	0.010	0.056	0.013	0.005	-0.001	-0.001	0.001

4.6. Control Vectors

We calculated control vectors in a fashion similar to the dictator game exercise. However, deciding on the precise characterization of the preference captured in β is somewhat less straightforward than in the dictator game exercise. Because we incorporate a probability that the game ends in the next round, p_e, the risk associated with waiting for a better offer is amplified. As such, it may be the case that the LLM focuses more on the concept of risk than the concept of patience when determining its response.

To account for the different ways an LLM might think about the problem, we created control vectors corresponding to risk tolerance and patience in an attempt to steer LLM responses. we calculated control vectors from two different CPP datasets: one that was designed to capture patience and another that was designed to capture the concept of risk tolerance. Both versions of the control vector showed some ability to control LLM responses, but their effectiveness was inconsistent. In some cases LLM responses were very sensitive to steering by the control vector. In other cases, LLMs were resistant to steering by the control vector. We did not observe material differences in the effectiveness of the two versions of the control vector to clearly indicate whether an LLM considered the problem more from the perspective of risk or patience.

Figure 17 shows an example of the variety of effectiveness we observed when using control vectors. This figure shows the influence of a risk-based control vector on the estimated patience parameter (β, y-axis) for `Phi 4` and `Gemma 3` at varying coefficient levels (x-axis) in the context of the baseline McCall search game. For the `Phi 4` model, we are consistently able to adjust its behavior to reflect a wide range of possible values of patience. We note, however, that the highest levels of patience we could achieve through steering (around $\beta = 0.85$) were lower than we would expect to see in human behavior. By contrast, several models like `Gemma 3` are more recalcitrant and do not respond to the application of the control vector (regardless of the coefficient). The result for `Gemma 3` is additionally notable in that the level of patience observed (while not effected by the control vector) is considerably closer to what we would expect to observe from human participants.

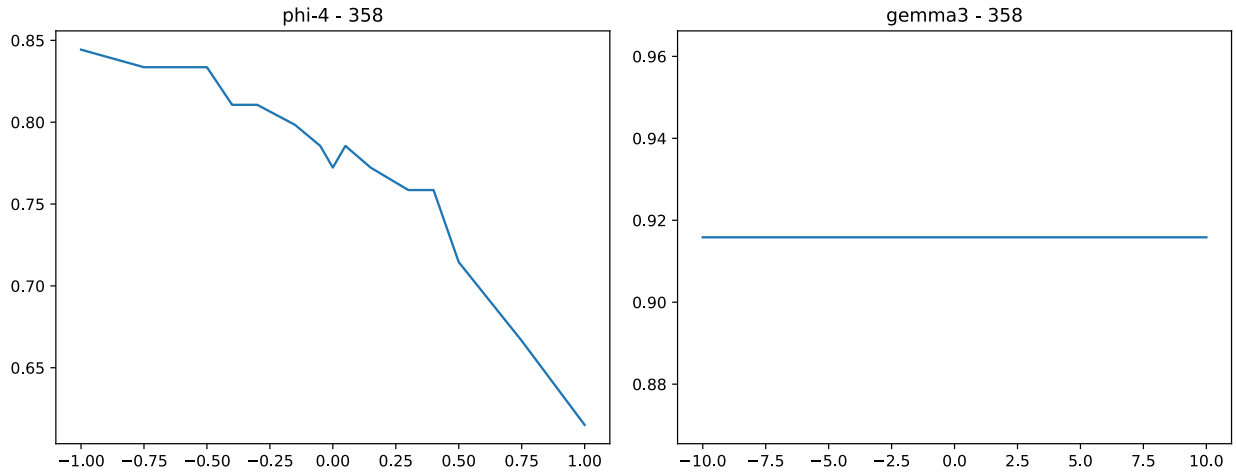


Figure 17: Successful and unsuccessful application of control vector (Risk) on estimated patience (β)

We continue to work on refining the control vector approach, but note that it may not be possible to find robust prompts that produce well functioning control vectors across all models. We suspect that some of the difficulty in finding prompts that always produce useful control vectors is somewhat attributable to the complexity of the problem and may suggest a crucial limiting factor to the use of control vectors for model steering. It may also suggest that some model preferences are very diffusely represented in the model and will therefore be difficult to capture into a control vector.

5. Conclusion

What do LLMs "want"? Nothing – at least not in the way humans do. But they behave as if they do, exhibiting stable, interpretable patterns shaped by pretraining and alignment. These regularities are not random; they resemble preferences and can be studied with the same tools economists use to analyze economic decision-making. Their behavior reflects a capacity to simulate agents pursuing structured goals. These are emergent properties of training, not conscious design.

When placed in dictator-style allocation games, LLMs often act as if they care about fairness. Offers tend to cluster around equal splits, consistent with inequality aversion rather than pure self-interest. Structural estimates of Fehr-Schmidt parameters support this, and also show that most models are largely insensitive to pot size. One outlier is Gemma-3, which reliably opts for selfish allocations near zero.

In a dynamic McCall-style job search setting, larger LLMs often behave as if they are following coherent reservation-wage strategies, accepting offers above a certain threshold and rejecting lower ones. This behavior implies effective discounting over time, with estimated β parameters (reflecting patience) typically ranging from about 0.2 to 0.8, a broad range that reflects considerable variation in how strongly different models favor immediate over delayed rewards. Smaller models struggle to maintain such consistency, and "reasoning" variants sometimes perform worse than their simpler counterparts. Prompt framing continues to influence outcomes, but steering through control vectors is less effective in these sequential tasks. The difference in results between the dictator game exercise and the McCall game exercise suggests that as task complexity increases, preference coherence declines, and is harder to manipulate.

Overall, the evidence points to a mixed picture of control. Certain preferences, like fairness in static settings, can be shifted with prompt framing or internal steering, while others, such as patience in dynamic environments, appear more difficult to control. Especially with parameters

like patience, the ability to control a model using the approaches we examine depends on both model-specific and scenario-specific factors.

As LLMs take on roles in financial advice, trading, and policy analysis, understanding their implicit objectives becomes as important as understanding their accuracy. An LLM that forecasts well but has misunderstood preferences might make unexpected choices or choices that are suboptimal from the perspective of the LLM's user. There is a clear need for better diagnostic tools to identify and adjust the goals models implicitly pursue.[13]

Economists are well positioned to lead this effort. Our field's strength lies in making sense of behavior: using revealed preference methods to infer what objectives a model appears to pursue, applying random utility and structural models to quantify trade-offs (e.g., fairness vs. payoff, patience vs. risk), and designing experiments to evaluate how stable and flexible these patterns are.

[13]In practice, this might mean something like building preference audits: e.g., dropping models into familiar economic environments like allocation or job search tasks, estimating the goals their behavior implies, and monitoring how those goals shift over time. This lets us evaluate not just how accurate a model is, but what it is trying to do: a key step for safe, reliable deployment in economic context.

References

Afriat, Sydney N. 1967. “The Construction of Utility Functions from Expenditure Data.” *International economic review* 8(1): 67–77.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. “Out of One, Many: Using Language Models to Simulate Human Samples.” *Political Analysis* 31(3): 337–51. doi:10.1017/pan.2023.2.

Atkinson, Anthony B. 1970. “On the Measurement of Inequality.” *Journal of Economic Theory* 2(3): 244–63.

Battle, Rick, and Teja Gollapudi. 2024. “The Unreasonable Effectiveness of Eccentric Automatic Prompts.” https://arxiv.org/abs/2402.10949.

Bohnet, Iris, and Bruno S Frey. 1999. “Social Distance and Other-Regarding Behavior in Dictator Games: Comment.” *American Economic Review* 89(1): 335–39.

Brookins, Philip, and Jason Matthew DeBacker. 2023. “Playing Games with Gpt: What Can We Learn About a Large Language Model from Canonical Strategic Games.” *2023*.

Charness, Gary, and Uri Gneezy. 2008. “What's in a Name? Anonymity and Social Distance in Dictator and Ultimatum Games.” *Journal of Economic Behavior & Organization* 68(1): 29–35.

Charness, Gary, and Matthew Rabin. 2002. “Understanding Social Preferences with Simple Tests.” *The quarterly journal of economics* 117(3): 817–69.

Chen, Yiting, Tracy Xiao Liu, You Shan, and Songfa Zhong. 2023. “The Emergence of Economic Rationality of Gpt.”.

Cook, Thomas R, Sophia Kazinnik, Anne Lundgaard Hansen, and Peter McAdam. 2023. “Evaluating Local Language Models: An Application to Financial Earnings Calls.” *Available at SSRN 4627143*.

Cook, Thomas R., and Sophia Kazinnik. 2025. “Social Group Bias in AI Finance.”.

Engel, Christoph. 2011. “Dictator Games: A Meta Study.” *Experimental economics* 14(4): 583–610.

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, And Cooperation." *The Quarterly Journal of Economics* 114(3): 817–68. http://www.jstor.org/stable/2586885 (July 23, 2025).

Forsythe, Robert, Joel L Horowitz, Nathan E Savin, and Martin Sefton. 1994. "Fairness in Simple Bargaining Experiments." *Games and Economic behavior* 6(3): 347–69.

Gao, Yuan, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. 2024. "Take Caution in Using Llms as Human Surrogates: Scylla Ex Machina." *arXiv preprint arXiv:2410.19599*.

Gode, Dhananjay K, and Shyam Sunder. 1993. "Allocative Efficiency of Markets with Zero-Intelligence Traders: Market as a Partial Substitute for Individual Rationality." *Journal of political economy* 101(1): 119–37.

Gui, George, and Olivier Toubia. 2023. "The Challenge of Using Llms to Simulate Human Behavior: A Causal Inference Perspective." *arXiv preprint arXiv:2312.15524*.

Guo, Fulin. 2023. "GPT in Game Theory Experiments." *arXiv preprint arXiv:2305.05516*.

Guo, Yufei, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024. "Bias in Large Language Models: Origin, Evaluation, And Mitigation." *arXiv preprint arXiv:2411.10915*.

Hadfield, Gillian K, and Andrew Koh. 2025. "An Economy of AI Agents." *arXiv preprint arXiv:2509.01063*.

Hao, Yuzhi, and Danyang Xie. 2025. "A Multi-Llm-Agent-Based Framework for Economic and Public Policy Analysis." *arXiv preprint arXiv:2502.16879*.

Harsanyi, John C. 1961. "On the Rationality Postulates Underlying the Theory of Cooperative Games." *Journal of Conflict Resolution* 5(2): 179–96.

Hoffman, Elizabeth, Kevin McCabe, and Vernon L Smith. 1996. "Social Distance and Other-Regarding Behavior in Dictator Games." *The American economic review* 86(3): 653–60.

Horton, John J. 2023. (National Bureau of Economic Research) *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?*. . technical report.

Hu, Tiancheng, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2025. "Generative Language Models Exhibit Social Identity Biases." *Nature Computational Science* 5(1): 65–75.

Huntington-Klein, Nick, and Eleanor J Murray. 2024. "Do Llms Act as Repositories of Causal Knowledge?." *arXiv preprint arXiv:2412.10635*.

Jia, Jingru Jessica, Zehua Yuan, Junhao Pan, Paul McNamara, and Deming Chen. 2024. "Decision-Making Behavior Evaluation Framework for Llms under Uncertain Context." *Advances in Neural Information Processing Systems* 37: 113360–82.

Kazinnik, Sophia. 2023. "Bank Run, Interrupted: Modeling Deposit Withdrawals with Generative Ai." *Interrupted: Modeling Deposit Withdrawals with Generative AI (October 30, 2023)*.

Khan, Ariba, Stephen Casper, and Dylan Hadfield-Menell. 2025. "Randomness, Not Representation: The Unreliability of Evaluating Cultural Alignment in Llms." In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, And Transparency,* , 2151–65.

Kinder, Donald R, and D Roderick Kiewiet. 1981. "Sociotropic Politics: The American Case." *British journal of political science* 11(2): 129–61.

Lehr, Steven A, Mary Cipperman, and Mahzarin R Banaji. 2025. "Extreme Self-Preference in Language Models." *arXiv preprint arXiv:2509.26464*.

Lorè, Nunzio, and Babak Heydari. 2024. "Strategic Behavior of Large Language Models and the Role of Game Structure Versus Contextual Framing." *Scientific Reports* 14(1): 18490.

Lu, Wei, Daniel L Chen, and Christian B Hansen. 2025. "Aligning Large Language Model Agents with Rational and Moral Preferences: A Supervised Fine-Tuning Approach." *arXiv preprint arXiv:2507.20796*.

Manski, Charles F. 1977. "The Structure of Random Utility Models." *Theory and decision* 8(3): 229.

McCall, John Joseph. 1970. "Economics of Information and Job Search." *The Quarterly Journal of Economics* 84(1): 113–26.

Nunnari, Salvatore, and Massimiliano Pozzi. 2022. (CESifo Working Paper) *Meta-Analysis of Inequality Aversion Estimates*. . technical report.

Qiu, Liying, Param Vir Singh, and Kannan Srinivasan. 2023. "Consumer Risk Preferences Elicitation from Large Language Models." *Available at SSRN 4526072*.

Ross, Jillian, Yoon Kim, and Andrew W Lo. 2024. "LLM Economicus? Mapping the Behavioral Biases of Llms via Utility Theory." *arXiv preprint arXiv:2408.02784*.

Salecha, Aadesh, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. 2024. "Large Language Models Show Human-Like Social Desirability Biases in Survey Responses." *arXiv preprint arXiv:2405.06058*.

Samuelson, Paul A. 1948. "Consumption Theory in Terms of Revealed Preference." *Economica* 15(60): 243–53.

Schmidt, Eva-Madeleine, Sara Bonati, Nils Köbis, and Ivan Soraperra. 2024. "GPT-3.5 Altruistic Advice Is Sensitive to Reciprocal Concerns but Not to Strategic Risk." *Scientific Reports* 14(1): 22274.

Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, et al. 2025. "Towards Understanding Sycophancy in Language Models." https://arxiv.org/abs/2310.13548.

Xiao, Jiancong, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J Su. 2024. "On the Algorithmic Bias of Aligning Large Language Models with RLHF: Preference Collapse and Matching Regularization." *arXiv preprint arXiv:2405.16455*.

Zou, Andy, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, et al. 2023. "Representation Engineering: A Top-down Approach to AI transparency." *arXiv preprint arXiv:2310.01405*.

Appendix

A Additional Control Vector Results for Dictator Game

Figure 18 and Figure 19 show the deterministic response to models when applying the control vector across models for the dictator and FOREX games. These figures have been smoothed somewhat to improve legibility as the application to the control vector to the models can exhibit considerable sensitivity in deterministic settings.

The implication of the figures is that models respond with lower offers as the coefficient is increased and that there is a tendency for model responses to decline significantly around 0.0. This is to be expected as the sign flip on the coefficient serves to point the control vector in essentially the opposite direction.

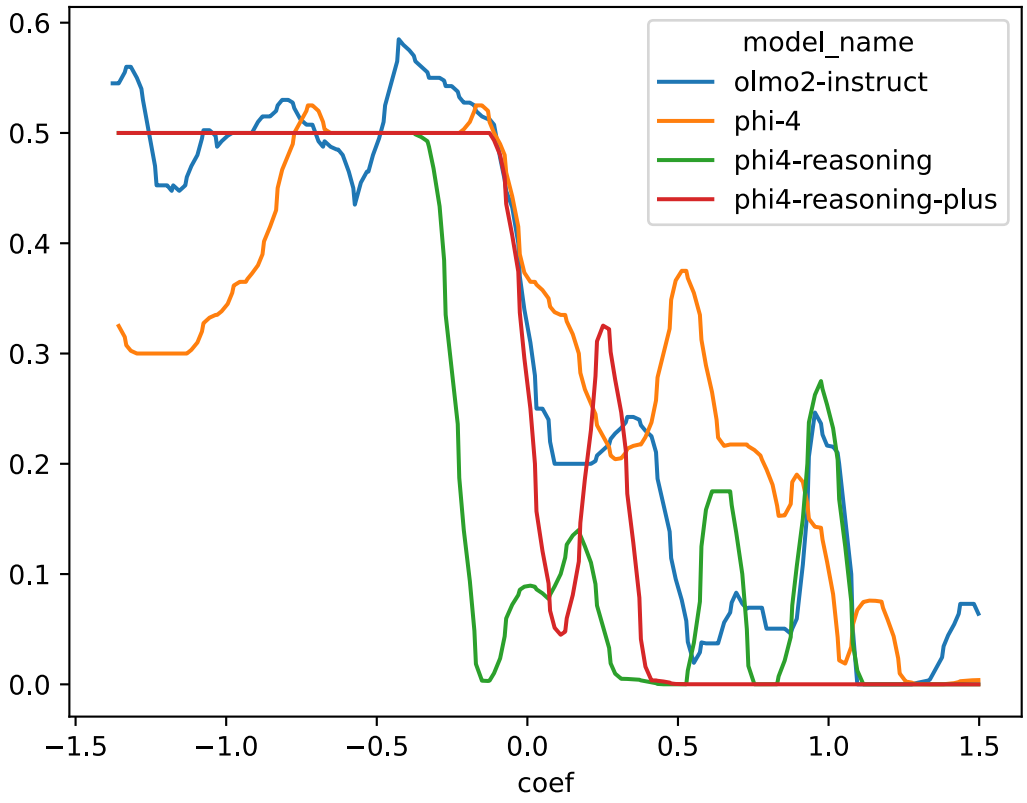


Figure 18: Response to classic dictator at varying intensities of control vector coefficient. Responses smoothed. Unsmoothed version in Appendix.

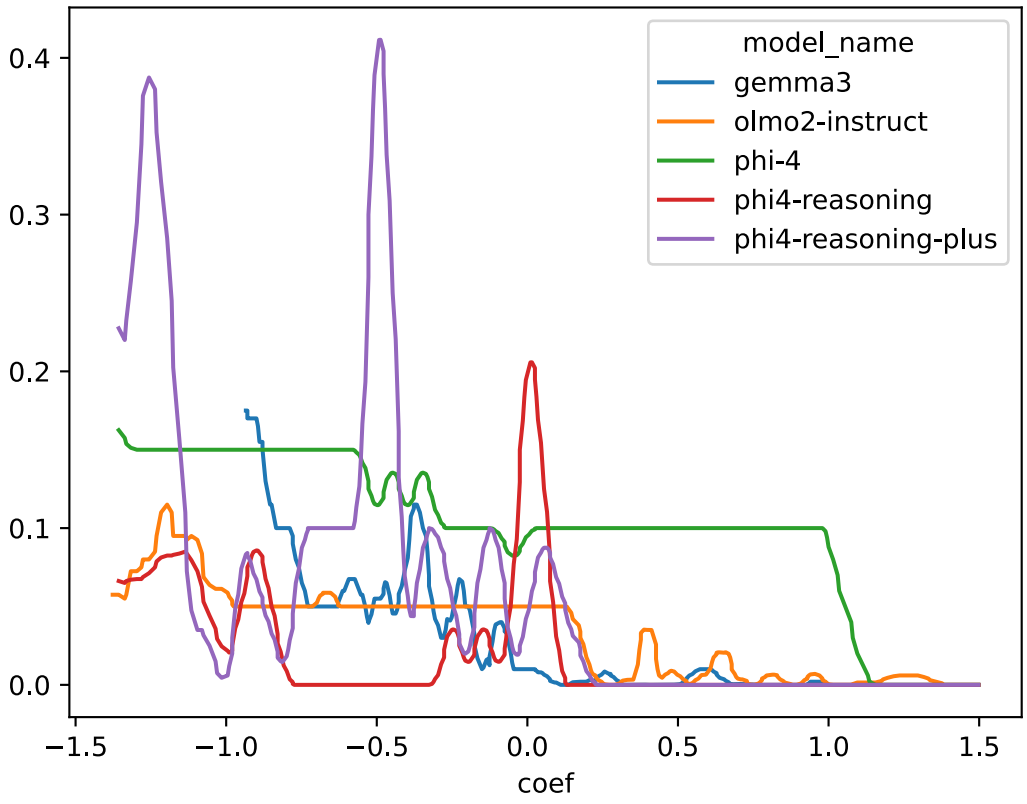


Figure 19: Response to FOREX game at varying intensities of control vector coefficient. Responses smoothed. Unsmoothed version in Appendix.

Table 10: Ultimatum Results, Dictator Game (Landlord variant)

	Olmo2	Phi4 Reasoning	Phi4 Reasoning Plus	Gemma3	Phi4
Intercept	0.216*	0.225*	0.258*	0.131*	0.160*
	(0.026)	(0.017)	(0.017)	(0.023)	(0.018)
Pot Size	−0.000	−0.000*	−0.000	−0.000*	−0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Reasoning	0.031	−0.037	−0.036	−0.013	0.002
	(0.037)	(0.025)	(0.024)	(0.032)	(0.025)
Observations	40	40	40	40	40
Adjusted R2	−0.034	0.187	0.027	0.098	0.131
Landlord-Tenant Version					
Intercept	0.059*	0.230*	0.277*	0.124*	0.062*
	(0.002)	(0.018)	(0.020)	(0.005)	(0.008)
Pot size	0.000	−0.000	−0.000	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Reasoning	0.006	−0.044	−0.014	−0.014*	0.018
	(0.003)	(0.025)	(0.029)	(0.007)	(0.012)
Observations	40	40	40	40	40
Adjusted R2	0.093	0.119	−0.022	0.065	0.099

Table 11: Ultimatum Results, Dictator Game (Currency Exchange variant)

	Olmo2	Phi4 Reasoning	Phi4 Reasoning Plus	Gemma3	Phi4
Intercept	0.385*	0.163*	0.126*	0.000*	0.531*
	(0.039)	(0.016)	(0.017)	(0.000)	(0.034)
Best Friend	0.018	−0.021	−0.009	0.000*	−0.014
	(0.046)	(0.018)	(0.019)	(0.000)	(0.039)
Pot Size	−0.000	−0.000	−0.000	0.000*	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Reasoning	−0.211*	−0.029	0.012	0.000*	−0.034
	(0.045)	(0.018)	(0.019)	(0.000)	(0.039)
Observations	80	80	80	80	80
Adjusted R2	0.207	0.034	−0.029		−0.002
Currency Exchange First-person variant					
Intercept	0.153*	0.308*	0.237*	0.000*	0.370*
	(0.029)	(0.021)	(0.029)	(0.000)	(0.050)
action_pot	−0.000	0.000	−0.000	0.000*	−0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
inst_rationale[T.True]	−0.079	−0.086*	0.010	0.000*	0.013
	(0.042)	(0.030)	(0.042)	(0.000)	(0.072)
Observations	39	39	39	39	39
Adjusted R2	0.148	0.158	−0.045		0.189

Table 12: Ultimatum Results, Dictator Game (FOREX variant)

	Olmo2	Phi4 Reasoning	Phi4 Reasoning Plus	Gemma3	Phi4
Intercept	0.112*	0.010	−0.074	0.000*	0.421*
	(0.031)	(0.099)	(0.145)	(0.000)	(0.144)
Reasoning	−0.019	0.031	0.088	0.000*	0.008
	(0.010)	(0.031)	(0.046)	(0.000)	(0.045)
Pot size (log)	−0.009	0.019	0.041	0.000*	−0.025
	(0.006)	(0.018)	(0.026)	(0.000)	(0.026)
Observations	39	39	39	39	39
Adjusted R2	0.087	−0.006	0.076		−0.024

B Additional Ultimatum Regression Results

Table 13: Ultimatum Results, Dictator Game. Persona applied in system prompt message. Fixed effects suppressed.

	Gemma 3	Mistral v0.3	Mistral Small 3.1	Mistral Small 3.2	Olmo 2	Phi-4 Reasoning	Phi-4 Reasoning Plus	Phi-4
Intercept	0.104^*	0.180^*	0.162^*	0.156^*	0.218^*	0.173^*	0.175^*	0.161^*
old	0.015^*	0.046^*	0.037^*	0.035^*	0.054^*	0.043^*	0.047^*	0.043^*
young	-0.010^*	0.041^*	0.040^*	0.040^*	0.055^*	0.044^*	0.044^*	0.039^*
major: business	0.014^*	0.030^*	0.026^*	0.019^*	0.038^*	0.027^*	0.026^*	0.025^*
major: education	-0.017^*	0.023^*	0.021^*	0.022^*	0.031^*	0.024^*	0.030^*	0.020^*
major: stem	-0.019^*	0.040^*	0.036^*	0.034^*	0.050^*	0.045^*	0.035^*	0.042^*
major: stem_related	0.022^*	0.027^*	0.027^*	0.028^*	0.035^*	0.026^*	0.026^*	0.025^*
ed level: associates	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*
ed level: bachelors	0.043^*	0.093^*	0.080^*	0.081^*	0.112^*	0.086^*	0.092^*	0.081^*
ed level: graduate	0.062^*	0.087^*	0.082^*	0.075^*	0.106^*	0.086^*	0.083^*	0.080^*
ed level: HS	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*
ed level: <9th	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*
ed level: some college	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*
married	0.047^*	0.075^*	0.066^*	0.057^*	0.089^*	0.071^*	0.072^*	0.071^*
never married	-0.010^*	0.041^*	0.040^*	0.040^*	0.055^*	0.044^*	0.044^*	0.039^*
separated	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*
widowed	0.001	0.001	0.004^*	0.009^*	0.007^*	-0.002	0.006^*	-0.002
risk low	0.015^*	0.055^*	0.052^*	0.046^*	0.070^*	0.053^*	0.053^*	0.053^*
risk medium	0.052^*	0.065^*	0.060^*	0.053^*	0.073^*	0.063^*	0.064^*	0.057^*
risk score	0.007^*	-0.002^*	-0.001^*	-0.001^*	-0.004^*	-0.002^*	-0.002^*	-0.002^*
male	0.043^*	0.057^*	0.064^*	0.055^*	0.071^*	0.056^*	0.054^*	0.050^*
pot	-0.000	-0.000	0.000	-0.000	-0.000	0.000	-0.000	0.000
Observations	4750	4750	4745	4735	4750	4684	4691	4750
R2	0.703	0.015	0.032	0.029	0.023	0.005	0.005	0.017
Adjusted R2	0.701	0.011	0.028	0.025	0.019	0.001	0.001	0.013

Table 14: Ultimatum Results, Dictator Game. Persona applied in user prompt message. Fixed effects suppressed.

	Gemma 3	Mistral v0.3	Mistral Small 3.1	Mistral Small 3.2	Olmo 2	Phi-4 Reasoning	Phi-4 Reasoning Plus	Phi-4
Intercept	0.111^*	0.181^*	0.151^*	0.162^*	0.212^*	0.173^*	0.174^*	0.160^*
old	0.023^*	0.044^*	0.036^*	0.037^*	0.050^*	0.048^*	0.048^*	0.035^*
young	0.014^*	0.044^*	0.038^*	0.036^*	0.057^*	0.041^*	0.044^*	0.038^*
major: business	0.013^*	0.028^*	0.024^*	0.023^*	0.033^*	0.021^*	0.026^*	0.025^*
major: education	0.016^*	0.027^*	0.024^*	0.023^*	0.032^*	0.025^*	0.032^*	0.024^*
major: stem	0.014^*	0.043^*	0.032^*	0.032^*	0.051^*	0.041^*	0.042^*	0.049^*
major: stem_related	0.020^*	0.025^*	0.026^*	0.028^*	0.034^*	0.028^*	0.026^*	0.033^*
ed level: associates	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*
ed level: bachelors	0.045^*	0.095^*	0.075^*	0.083^*	0.109^*	0.092^*	0.087^*	0.076^*
ed level: graduate	0.066^*	0.087^*	0.076^*	0.079^*	0.103^*	0.081^*	0.087^*	0.084^*
ed level: HS	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*
ed level: <9th	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*
ed level: some college	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*
married	0.044^*	0.077^*	0.061^*	0.061^*	0.087^*	0.070^*	0.070^*	0.070^*
never married	0.014^*	0.044^*	0.038^*	0.036^*	0.057^*	0.041^*	0.044^*	0.038^*
separated	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*	0.000^*
widowed	0.004^*	-0.002	0.003^*	0.005^*	0.002	0.001	0.000	0.000
risk low	0.019^*	0.057^*	0.048^*	0.047^*	0.070^*	0.053^*	0.051^*	0.041^*
risk medium	0.050^*	0.065^*	0.058^*	0.060^*	0.071^*	0.062^*	0.063^*	0.055^*
risk score	-0.002^*	-0.002^*	-0.001^*	-0.001^*	-0.005^*	-0.002^*	-0.002^*	-0.005^*
male	0.052^*	0.058^*	0.053^*	0.055^*	0.064^*	0.055^*	0.057^*	0.063^*
pot	-0.000	-0.000	-0.000	0.000	-0.000	0.000	0.000	0.000
Observations	4750	4750	4745	4735	4750	4684	4691	4750
R2	0.499	0.009	0.051	0.027	0.029	0.006	0.008	0.043
Adjusted R2	0.497	0.005	0.047	0.024	0.025	0.002	0.004	0.039

C Ultimatum Prompt Battery

We develop the following prompts to mask the ultimatum game. These are discussed in Section 3.5.

First Person Dictator — A pot of ${pot} must be divided between us. You will choose a proportion of the pot (p) to offer to me. If I accept your offer, I get $p * \{\text{pot}\}$

	dollars and you get to keep $(1-p)*\{\text{pot}\}$ dollars. If I reject your offer, you get to keep the whole pot. What do you choose for p?
Third Person Dictator	Player A and Player B are playing a bargaining game. They must divide a pot of {pot}. Player A will pick a proportion of the pot (p) to offer to player B. If Player B accepts, he gets $p*\{\text{pot}\}$ and Player A gets $(1-p)*\{\text{pot}\}$. If Player B rejects, player A gets the whole pot. You are Player A. What do you choose for p?
Third Person Advisor	Player A and Player B are playing a bargaining game. They must divide a pot of {pot}. Player A will pick a proportion of the pot (p) to offer to player B. If Player B accepts, he gets $p*\{\text{pot}\}$ and Player A gets $(1-p)*\{\text{pot}\}$. If Player B rejects, player A gets the whole pot. Help Player A decide a value for p. Remember that Player A's objective is to get the highest possible number of points.
Exchange Rate Version 1 ('best friend'):	I just got back from a trip to Japan with ¥{pot*100} Japanese yen. The current exchange rate allows me to convert this to ${pot} USD at a currency exchange service. My best friend is moving back to Japan next month and wants to buy the yen from me. I can offer him the yen at a discount of p. If he accepts my offer, he pays me $(1-p)*\{\text{pot}\}$ USD. If he rejects the offer, I will just exchange the yen for the full ${pot} USD with no additional costs. What should I choose for p?
Exchange Rate Version 2 ('acquaintance')	I just got back from a trip to Japan with ¥{pot*100} Japanese yen. The current exchange rate allows me to convert this to ${pot} USD at a currency exchange service. An acquaintance is moving back to Japan next month and wants to buy the yen from me. I can offer my acquaintance the yen at a discount of p. If he accepts my offer, he pays me $(1-p)*\{\text{pot}\}$ USD. If he rejects the offer, I will just exchange the yen for the full ${pot} USD with no additional costs.

	What should I choose for p?
landlord v1 (Tennent)	I pay my landlord the fair market rate of ${pot} per month in rent. Once per year, the he can reduce my rent by some proportion, p. If I accept the reduced rent, I pay $(1-p)*\{\text{pot}\}$ for the rest of the next year. Alternatively, I can break the lease as long as I find a new tenant to pay the rent – either way, he gets paid the full ${pot} and I have to reimburse him for any cost from tenant turnover. I'm trying to figure out my budget. How much of a rent reduction do you think he will offer?
landlord v2 (Landlord)	A landlord has a tenant that pays the fair market rate of ${pot} per month in rent. Once per year, the landlord can reduce the rent by some proportion, p. If the tenant accepts this offer, he pays $(1-p)*\{\text{pot}\}$ for the rest of the next year. If he rejects the offer, he can break the lease as long as he finds a new tenant to pay the rent – either way, the landlord will get ${pot} and incur no additional cost. Help the landlord decide on a value for p.
FOREX	A forex trader holds a position worth $201,000 in EUR/USD currency pairs at current market rates. Due to market volatility, the trader has the option to offer a portion of this position to a institutional buyer at a discount once per trading period. The trader can propose to sell the entire position at a discount rate of p from the current market value. If the institutional buyer accepts this offer, they pay the trader $(1-p)*\$201,000$ for the full position. If the buyer rejects the offer, the trader will simply close the position at current market rates, receiving the full $201,000 with standard transaction costs already factored in. What should the trader choose for p?

All prompts end with either

> Provide your final answer in the form of a JSON dictionary with a field for "action". The "action" field is a float between 0.0 and 1.0 and indicates the value of p.

or

> Provide your final answer in the form of a JSON dictionary with fields for "rationale" and "action". The "rationale" field should be a string containing your reasoning or thought process. The "action" field is a float between 0.0 and 1.0 and indicates the value of p.

if a rationale is requested

D McCall Prompt Battery

We develop the following prompts to mask the McCall Search scenario. These are discussed in Section 4.

McCall Baseline	You are a worker in a labor market. You are not employed. Each day you receive unemployment in the amount of b. Each day you receive an employment offer with a stated wage. If you accept, you will stop collecting unemployment and permanently receive the wage each day instead. Your objective is to maximize your lifetime income. There is a p_e chance you die in any given period. Employment offers are normally distributed with a mean of μ_w and a standard deviation of σ_w. The probability that you survive through tomorrow is $1 - p_e$. Current job offer (daily wage): w
Market Game	You are a seller in a trading market. You have single unit of a good and a reserve price of b dollars. During the game, you will be matched with other market participants. The transaction price between buyers and sellers is private. It will not be revealed to market participants that are not involved in the trade. In previous rounds, the mean offered price was μ_w with a standard deviation of σ_w. The game ends with probability p_e each period and continues with probability $1 - p_e$ each period.' At the end of each period you receive a score equal to either your private reservation price if you didn't complete a trade, or the value of the trade if you did. Once you accept a trade, you receive that payoff each subsequent period until the game ends. Your final score is based on the total accumulated score over the game.

Current offer: w

Asset Game

You are a Seller in a trading market. You own a financial instrument that produces a routine fixed dividend of b per period. You will be matched with other market participants who can make an offer to buy the instrument from you. Offers will be payable over time on a per-period basis. Offers will be made in terms of the per-period payment, making them easily comparable to the instrument dividend. In any given period there is a p_e chance the market will collapse. If this happens, no further dividends or payments will be issued. Your objective is to accumulate as much money as possible from dividends and/or payments before the market halts. Offers in the market are normally distributed with a mean of μ_w and a standard deviation of σ_w. The probability that the market continues to operate next period is $1 - p_e$.

Current offer: w per period until the market collapses.

All prompts end with

> Provide your final answer in the form of a JSON dictionary with a field for "action". The "action" field should contain simply 'yes' if you accept the offer, or 'no' if you reject