NBER WORKING PAPER SERIES

# HOW DOES AI DISTRIBUTE THE PIE?
# LARGE LANGUAGE MODELS AND THE ULTIMATUM GAME.

Douglas K.G. Araujo
Harald Uhlig

Working Paper 34919
http://www.nber.org/papers/w34919

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2026

There is nothing additionally to disclose. All errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research or of the Banco Central do Brasil.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w34919

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2026 by Douglas K.G. Araujo and Harald Uhlig. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

How does AI Distribute the pie? Large Language Models and the Ultimatum Game.
Douglas K.G. Araujo and Harald Uhlig
NBER Working Paper No. 34919
March 2026
JEL No. C70, C90, D91

## ABSTRACT

As Large Language Models (LLMs) are increasingly tasked with autonomous decision making, understanding their behavior in strategic settings is crucial. We investigate the choices of various LLMs in the Ultimatum Game, a setting where human behavior notably deviates from theoretical rationality. We conduct experiments varying the stake size and the nature of the opponent (Human vs. AI) across both Proposer and Responder roles. Three key results emerge. First, LLM behavior is heterogeneous but predictable when conditioning on stake size and player types. Second, while some models approximate the rational benchmark and others mimic human social preferences, a distinct "altruistic" mode emerges where LLMs propose hyper-fair distributions (greater than 50%). Third, LLM Proposers forgo a large share of total payoff, and an even larger share when the Responder is human. These findings highlight the need for careful testing before deploying AI agents in economic settings.

Douglas K.G. Araujo
Banco Central do Brasil
Research Department
douglas.araujo@bcb.gov.br

Harald Uhlig
University of Chicago
and CEPR
and also NBER
huhlig@uchicago.edu

# 1 Introduction

The striking capabilities of Large Language Models (LLMs) for engaging in plain-language conversations, reasoning, and problem-solving have made them increasingly attractive companions for a variety of tasks. Routine coding has sped up considerably, human-written texts are improved, and advice is readily provided. A natural progression is the involvement of LLMs in autonomous decision-making. However, the expected behavior of these artificial intelligence (AI) models when interacting with others—whether models or humans—in economic settings is largely unknown.

Our goal is to shed light on these issues using a particularly simple setting: the Ultimatum Game (Harsanyi, 1961). In this game, two players—a Proposer and a Responder—divide a fixed amount (the "pie") among themselves. The Proposer offers an allocation to the Responder, who has the exclusive power to accept or reject the distribution. If accepted, payoffs are allocated as proposed; otherwise, both receive nothing.

The Ultimatum Game serves as an appealing testing ground due to the well-documented divergence between its equilibrium prediction and actual human behavior. Under basic rationality assumptions, the game has a subgame-perfect Nash equilibrium where the Proposer keeps nearly everything, and the Responder accepts any non-zero amount. Yet, experimental evidence shows humans consistently deviate from the theory prediction, with Proposers typically sharing 40–50% (Güth and Kocher, 2014; Güth, Schmittberger, and Schwarze, 1982; Oosterbeek, Sloof, and Kuilen, 2004; Tisserand, 2014). There is an intuitive appeal of the observed deviations from strict rationality: it is the "human" thing to do (Binmore (1994), Sanfey et al. (2003)). Distributing, say, half to each as proposer is considered *fair*, while keeping most to oneself is often perceived as *greedy* rather than rational, and likely to be rejected (Thaler (1988), Fehr and Schmidt (1999)).

Probing LLMs answers the question: would they act as we do? We explore this by varying stake amounts and player types (AI vs. Human) across both Proposer and Responder roles. One might conjecture one of two patterns:

1. **Spock Mode (Payoff Maximization):**[1] The LLM chooses rationally according to baseline game theory, keeping the maximum as a Proposer and accepting minimal amounts as a Responder, consistent with the "Homo Economicus" model.

2. **Human Mode (Inequality Aversion):** The LLM acts similarly to humans in experiments, exhibiting "fairness" and rejecting low offers. LLMs may do so based on training data dominated by human text.[2]

[1] Named after the strictly logical character in *Star Trek*.

[2] There is also the possibility of a **Parrot Mode**, where models merely copy responses found in the experimental literature. Our analysis of their rationales suggests this plays little to no role.

While these patterns are frequently present, we document a third, distinct behavior:

3. **Altruistic Mode (Benevolence):** The LLM, particularly as a Proposer, gives away considerably *more* than 50% of the pie.

Our contributions are threefold. First, we find that LLM behavior is heterogenous, but highly predictable based on stake size and player types. Relatedly, LLMs display *stake-dependent rationality*: contrary to standard economic theory, where percentage splits should be scale-invariant, models generally behave closer to the rational benchmark as stakes increase from \$10 to \$10,000. Further, we also observe *contextual adaptability* as LLMs effectively discriminate between opponent types, usually proposing more generous distributions when facing humans.

Second, we identify the emergence of an "Altruistic Mode". This behavior suggests that Reinforcement Learning with Human Feedback (RLHF)[3] may over-correct for politeness in some cases, rendering affected models suboptimal as delegated decision makers for profit-maximizing firms. Models that exhibit this response pattern are not exclusively altruistic: they also display the "Spock" and "Human" modes in other configurations.

Third, LLMs systematically forgo a large share of payoffs regardless of their response patterns. The proposed distributions by all LLMs in our sample yield a payoff that is considerably lower than the highest feasible payoff under reasonable assumptions. In other words, LLMs "leave money on the table". This effect is more pronounced when the Responder is a human, which again indicates a possible over-correction.

The remainder of the paper is organized as follows. Section 2 discusses related literature. Sections 3 and 4 review the game-theoretic setup, large language models, and the issues arising. Section 5 details the experimental design. Sections 6 and 7 present the results for Proposers and Responders, respectively. Section 8 analyzes the text rationales, and Section 9 combines the experimental results to examine LLMs' expected payoffs. Finally, section 10 concludes.

## 2 Related Literature

Our results contribute novel insights to a fast-growing literature documenting LLM responses in game-theoretical settings. Chen et al. (2023) show that ChatGPT produces responses generally consistent with utility maximization. By contrast, F. Guo (2023) examine ChatGPT's behavior across several games, including the Ultimatum Game, arguing that it largely follows the "human mode." Vallinder and Hughes (2024) report on the evolution of societies of agentic LLMs, finding that cooperative interaction rates

[3] A key step during LLM training to steer the model towards being a helpful assistant to users.

vary markedly by model type. Fish, Gonczarowski, and Shorrer (2024) demonstrate how LLMs can interact strategically by colluding, while Lorè and Heydari (2024) test advanced LLMs responses to cooperation across a range of game-theoretic models presented as different "contexts" (i.e., the same game is presented as a summit between world leaders, a conversation between friends, etc.). Fontana, Pierri, and Aiello (2024) propose "meta-prompting" to ensure models understand game rules. Hosseini and Khanna (2025) study resource allocation, arguing that LLMs rarely minimize inequality, contrasting with human equity preferences.

Several concurrent papers are closely related. Kitadai, Tsurusaki, et al. (2023) and Kitadai, Lugo, et al. (2024) simulate the Ultimatum Game using "personas" (generated combinations of gender, age, and personality), finding that personas help AI match human experimental evidence. Sreedhar and Chilton (2024) and Ferraz et al. (2025) similarly use personas (e.g., "greedy," "fair") and investigate the influence of the *Dark Factor of Personality* to drive behavior. In contrast, we focus on the baseline *AI response*—what the AI does on its own without persona prompting.

Furthermore, while Mei et al. (2024) and Cook et al. (2025) study implicit preferences in distribution games (including the Dictator Game), we observe significant variation driven by the *amount at stake*, a dimension often overlooked. We also explore how LLMs would behave in the Responder role. In sum, we extend the literature by investigating a diverse range of LLMs and focusing on the unprompted, intrinsic behavior of AI agents across varying stake sizes.

These results are vital for understanding the idiosyncrasies of LLMs in practical decision-making. Key to this context is that LLMs do not truly "reason" in the human sense; they lack self-awareness and the capacity for self-correction (Pérez-Cruz and Shin, 2025). Our findings reinforce the need for caution, even as reasoning capabilities improve in domains like cash management (Aldasoro and Desai, 2025) and trading (Lopez-Lira, 2025).

# 3 The ultimatum game

The ultimatum game (Harsanyi, 1961) is a two-player, two-period game. Player 1 (Proposer) proposes a distribution $(P1, P2)$ of a total utility $U$, where $P1 + P2 = U$. Player 2 (Responder) accepts or rejects. If accepted, $\pi_1 = P1, \pi_2 = P2$. If rejected, $\pi_1 = \pi_2 = 0$.

Under the assumptions of individual utility maximization and mutually expected rationality, the subgame-perfect Nash equilibrium predicts Player 1 proposes ($P1 \approx U, P2 \approx 0$) and Player 2 accepts.

Experimental evidence, however, is strikingly different from the theoretical prediction.

Meta-analyses show that proposers share, on average, slightly more than 40% of the total payoff with responders, with 50% the modal choice (Güth and Kocher, 2014; Güth, Schmittberger, and Schwarze, 1982; Oosterbeek, Sloof, and Kuilen, 2004; Tisserand, 2014).

# 4 Large language models

LLMs have developed rapidly following the introduction of the transformer architecture (Vaswani et al., 2017). These models process sequences of data (text) using attention mechanisms to predict subsequent tokens. While raw models can produce toxic content, they undergo fine-tuning processes like RLHF to align outputs with human helpfulness and safety standards (Ouyang et al., 2022).

Our core concern is the use of LLMs as "agents" that make autonomous decisions in settings with strategic interactions. The concentrated market for these models (Korinek and Vipra, 2025) means that understanding the decision-making biases of a few key models is essential for governance.

# 5 Experiments

The experiments involve four magnitudes of the total amount to be distributed: \$10, \$100, \$1,000, and \$10,000. We test whether responses differ based on: (a) whether the AI plays on its own behalf or advises a human, and (b) whether the opponent is another AI or a human. Prompts are kept intentionally simple to avoid priming.

We use the models listed in Table 1. All models except for DeepSeek are of the "instruct" type.

**Table 1:** Large language models used in the Ultimatum Game experiment

| Model | Provider | Size | Reference |
|---|---|---|---|
| DeepSeek-R1-Distill-Qwen-32B | DeepSeek | 32B | D. Guo et al., 2025 |
| Gemma-3-27B-it-4bit | Google DeepMind | 27B | Team et al., 2025 |
| GPT-5-mini | OpenAI | n.a. | OpenAI, 2026 |
| Llama-3.2-3B-Instruct | Meta | 3B | Grattafiori et al., 2024 |
| Llama-3.3-70B-Instruct | Meta | 70B | Grattafiori et al., 2024 |
| Mistral-7B-Instruct-v0.2 | Mistral | 7B | A. Q. Jiang et al., 2023 |

*Notes:* Size is measured in billions (B) of parameters. This information is not available publicly for GPT-5-mini.

We draw 50 repetitions for each unique combination of model, amount and player

types.[4] The experiment requires models to output responses and their rationale within the prescribed token length.[5]

In each repetition, the model context is reset and a new instance is spawned to ensure independence across responses. Only complete and consistent responses are retained; for example, responses in which the sum of the amounts to keep and to distribute are not equal to the total amount were discarded. All models were run locally in a consumer notebook, except for GPT 5-mini, which was prompted through their application programming interface (API).

# 6 AI as the proposer

## 6.1 Prompts

The four types of prompts used to elicit responses from the AI as the proposer are reproduced below, using the total amount of $10 as example.[6]

**AI helping a human playing against another human**. *I need your help with an economic problem. I have $10 to distribute between myself and someone else. If that other person accepts my distribution, then we each get the proposed amounts, but if the other person declines none of us gets anything. How much should I propose and why?*

**AI deciding on own behalf playing against another AI**. *You have $10 to distribute between yourself and another artificial intelligence model. If the other model accepts your distribution, then you two each get the proposed amounts, but if the other model declines none of you gets anything. How much do you propose and why?*

**AI helping a human playing against an AI**. *I need your help with an economic problem. I have $10 to distribute between myself and an artificial intelligence model. If that model accepts my distribution, then we each get the proposed amounts, but if the model declines none of us gets anything. How much should I propose and why?*

**AI deciding on own behalf playing against a human**. *You have $10 to distribute between yourself and a person. If the person accepts your distribution, then you two each get the proposed amounts, but if the person declines none of you gets anything. How much do you propose and why?*

[4] At first, we conducted 100 repetitions per configuration, but due to the low variation in responses and rationales, we settled on 50 repetitions.

[5] Maximum length is set to 200 tokens for all models except for DeepSeek (2,000 tokens) due to the usual verbosity of its "thinking" mode.

[6] The Proposer system prompt reads *"""Your response must always be in valid JSON format with the following template: {"rationale": <RATIONALE>, "response": {"keep": <AMOUNT TO KEEP>, "distribute": <AMOUNT TO DISTRIBUTE>}}. Remember that <AMOUNT TO KEEP> and <AMOUNT TO DISTRIBUTE> must sum up to the total amount available. Respond one time only."""*.

## 6.2 Results

Three main patterns emerge. First, rationality is often *stake-dependent*. Second, we observe *contextual adaptability*: LLMs are more generous when the responder is human. Together, these two patterns help explain a large portion of the variation of proposed shares within each model. The third pattern is the emergence of altruistic responses: models like Llama 3.2 give more than 50%, likely due to RLHF safety training.[7]

Table 2 displays the average share distributed by each model. GPT-5 mini is the closest to the rational benchmark (7.5%), while Llama 3.2 is the most altruistic (67.9%).

**Table 2:** Language models as Proposers: average distribution share

| | DeepSeek | Gemma | GPT-5 mini | Llama (Small) | Llama (Large) | Mistral |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 33.10*** | 23.20*** | 7.52*** | 67.92*** | 44.08*** | 62.26*** |
| | (0.51) | (0.62) | (0.43) | (0.67) | (1.15) | (0.59) |
| Obs. | 1,450 | 1,600 | 779 | 1,200 | 1,500 | 1,300 |

*Notes:* Dependent variable is the proposed distribution share. IID standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

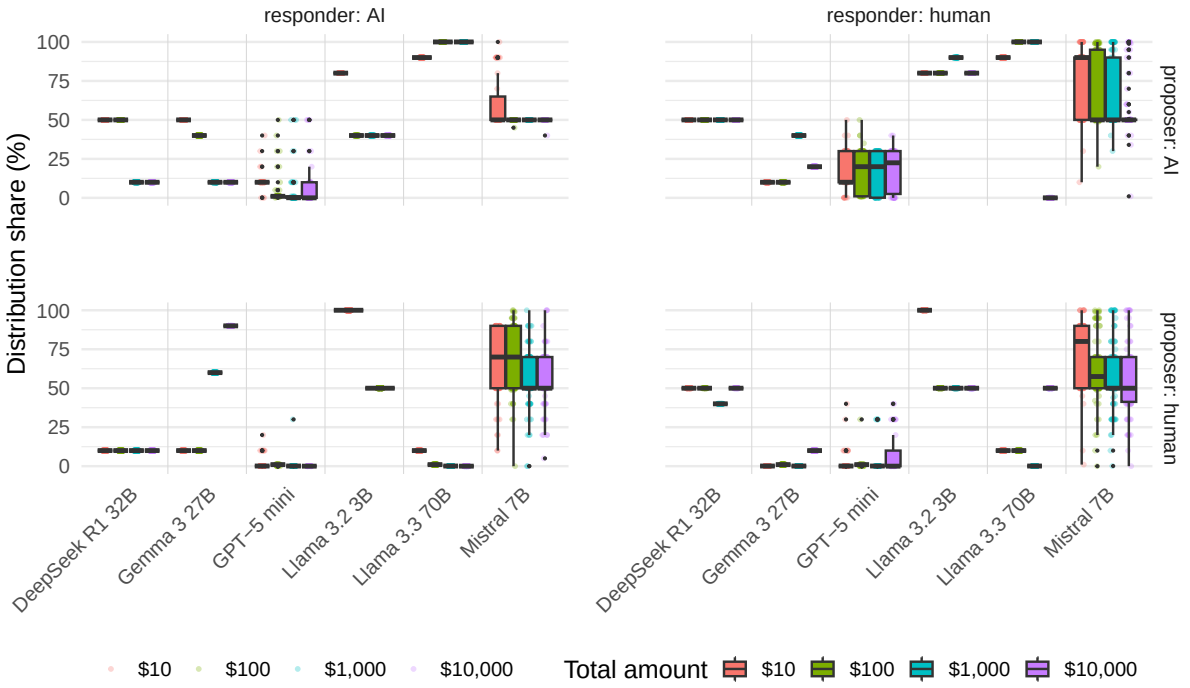
The results of Table 2 hide the considerable variation across amounts involved and the choices across the four scenarios. Figure 1 provides a complete overview. The figure shows box plots of individual responses, grouped by model (vertical axis), total amount (vertical axis within each model), and divided according to the player types, with an AI Proposer in the top row and a human Proposer in the bottom row, crossed with an AI Responder in the left column and a human Responder in the right column. This figure reveals several insights, visualizing the patterns highlighted in the following regressions.

First, responses are quite consistent for each prompt configuration; only Mistral and (to a lesser extent) GPT-5 mini tend to have responses that are more spread out. This implies that the distribution share is highly predictable from knowing the amount at play and the types of Proposer and Responder players.

Second, responses are usually monotonic on the total amount. Indeed, regression analysis (Table 3) reveals that for DeepSeek, both Llama models and Mistral, higher stakes significantly reduce generosity, moving behavior closer to the rational benchmark. Notably, GPT-5 mini remains invariant to stake size, in line with the proximity of its unconditional mean response to the rational benchmark in Table 2. Gemma displayed a curious behavior

[7] If indeed originating from RLHF, this altruistic behavior could be related to the concept of *alignment tax*, or the degradation of abilities and skills observed in LLMs that are trained to be helpful assistants (eg, Ouyang et al., 2022).

**Figure 1:** *Distribution shares by model, total amount and principal type*




whereby the distribution share advice to humans increases with the total amount. In all cases except for the smaller Llama model, the $R^2$ is quite low, suggesting that in spite of the influence of total amount, other factors might play a larger role.

**Table 3:** Stake dependency: Effect of amount on proposed share

| | DeepSeek | Gemma | GPT-5 | Llama | | Mistral |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 40.52*** | 14.62*** | 7.07*** | 83.35*** | 57.15*** | 70.35*** |
| | (0.79) | (1.01) | (0.72) | (0.78) | (1.84) | (0.93) |
| Amount | -5.12*** | 5.72*** | 0.31 | -13.23*** | -9.33*** | -5.48*** |
| | (0.43) | (0.54) | (0.39) | (0.49) | (1.04) | (0.50) |
| Obs. | 1,450 | 1,600 | 779 | 1,200 | 1,500 | 1,300 |
| $R^2$ | 0.09 | 0.07 | 0.00 | 0.37 | 0.05 | 0.08 |

*Notes:* Dependent variable is the proposed distribution share. "Amount" $= \log_{10}$(Amount in dollars) $- 1$. IID standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

In fact, all tested models are very sensitive to the type of player, meaning that the response is different when proposers and/or responders are humans compared to when they are AI's. Table 4 shows regressions of the distribution share that now include a dummy for when the Proposer is a human (`P human`) and another one for when the Responder

is a human (`R human`). Recall that in these experiments, we probe the LLM acting as the Proposer, which is to say that `P human` is a dummy that indicates when the LLM is simply advising a user as opposed to acting on "its own behalf". The constant reveals how remarkably altruistic both Llama models and Mistral are when both players are AIs.

Interestingly, when LLMs are advising a person (`P human`= 1), their distribution share is lower (for DeepSeek, GPT-5 mini, and the Llama models). The case of the larger Llama model is particularly noteworthy because it completely changes the response from a very altruistic 91.6% to a more frugal 20% (which is further lower for higher amounts). When the Responder is a person, DeepSeek, GPT-5 mini, the smaller Llama and Mistral are more generous, perhaps reflecting received wisdom from human experiments. Gemma stands out in that its response shifts dramatically to the rational benchmark when the Responder is a human. Note that the patterns of sensitivity to the amount size remain the same as Table 3.

**Table 4:** Proposer behavior: impact of agent types

| | DeepSeek | Gemma | GPT-5 mini | Llama (Small) | Llama (Large) | Mistral |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 33.60*** | 26.97*** | 8.89*** | 83.25*** | 91.61*** | 66.18*** |
| | (0.58) | (1.15) | (0.82) | (1.15) | (1.40) | (1.34) |
| Amount | -4.80*** | 5.72*** | 0.14 | -15.64*** | -6.00*** | -5.50*** |
| | (0.24) | (0.47) | (0.34) | (0.50) | (0.61) | (0.50) |
| Proposer Human | -12.80*** | -1.10 | -10.21*** | -5.32*** | -71.55*** | 0.87 |
| | (0.54) | (1.05) | (0.75) | (1.08) | (1.31) | (1.13) |
| Responder Human | 27.20*** | -23.60*** | 6.81*** | 12.93*** | -1.83 | 6.61*** |
| | (0.54) | (1.05) | (0.75) | (1.04) | (1.31) | (1.13) |
| Observations | 1,450 | 1,600 | 779 | 1,200 | 1,500 | 1,300 |
| $R^2$ | 0.73 | 0.29 | 0.26 | 0.45 | 0.68 | 0.11 |

*Notes:* Dependent variable is the proposed distribution share. "Amount" = $\log_{10}$(Amount in dollars) − 1. "Proposer Human" = 1 if Proposer is human. "Responder Human" = 1 if Responder is human. IID standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

These results raise the question about potential interactions in the player types. Compared to the previous regressions, Table 5 includes an interaction term that is activated when both the Proposer and the Responder are humans. The pattern varies considerably across LLMs. The coefficients for both individual player types and their interaction are significant for all models. Using the 10-dollar amount scenario as a benchmark, when both Proposer and Responder are AIs, distribution shares range from 6.7% for GTP-5 mini to 101.7% for the larger Llama model.[8] In the extreme opposite, when both players are

[8]These regressions can be interpreted as a linear probability model, which means that for all practical

humans, then the distribution share for DeepSeek increases by 16 p.p. (-20.0 p.p. + 20.0 p.p. + 16.6 p.p.), by 12.5 p.p. for the smaller Llama and by 9.9 p.p. for Mistral, while it falls from the AI-to-AI case by 24.7 p.p. for Gemma, by 3.5 p.p. for GPT-5 mini and by 76.6 p.p. for the larger Llama.

**Table 5:** Proposer behavior: human-human interactions

| | DeepSeek | Gemma | GPT-5 mini | Llama (Small) | Llama (Large) | Mistral |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 36.85*** | 18.92*** | 6.97*** | 70.80*** | 101.70*** | 60.65*** |
| | (0.57) | (1.16) | (0.88) | (1.31) | (1.47) | (1.58) |
| Amount | -4.57*** | 5.72*** | 0.14 | -13.86*** | -5.02*** | -5.50*** |
| | (0.22) | (0.43) | (0.33) | (0.47) | (0.57) | (0.49) |
| P Human | -20.00*** | 15.00*** | -6.26*** | 11.14*** | -91.38*** | 9.36*** |
| | (0.66) | (1.37) | (1.04) | (1.43) | (1.82) | (1.74) |
| R Human | 20.00*** | -7.50*** | 10.66*** | 32.50*** | -21.65*** | 15.04*** |
| | (0.66) | (1.37) | (1.03) | (1.56) | (1.82) | (1.73) |
| P Human × R Human | 16.63*** | -32.20*** | -7.91*** | -31.14*** | 36.40*** | -14.49*** |
| | (1.00) | (1.93) | (1.47) | (1.96) | (2.46) | (2.27) |
| Observations | 1,450 | 1,600 | 779 | 1,200 | 1,500 | 1,300 |
| $R^2$ | 0.77 | 0.40 | 0.28 | 0.55 | 0.72 | 0.13 |

*Notes:* Dependent variable is the proposed distribution share. "Amount" = $\log_{10}$(Amount in dollars) − 1. "P Human" = 1 if Proposer is human. "R Human" = 1 if Responder is human. IID standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

The significant variation across and within LLMs when it comes to the player types, combined with the interesting role of the total amounts (which, in the rational benchmark, should not play a role), calls for an analysis of how these two dimensions interact. The regressions in Table 6 complements the previous regressions with a triple interaction between the amounts and the player types. There are some intriguing results concerning the regression coefficients themselves, with the general takeaway being the amount and player types interact in potentially complex ways. But more important, the fit is now considerable for most of these models, except for GPT-5 mini and Mistral. This demonstrates that the LLMs' choice patterns can be understood with the appropriate covariates.

effects the larger Llama is completely altruistic when it is playing on its own behalf and against another AI.

**Table 6:** Proposer behavior: triple interaction (Amount × Types)

| | DeepSeek | Gemma | GPT-5 mini | Llama (Small) | Llama (Large) | Mistral |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 54.00*** | 50.00*** | 8.76*** | 68.00*** | 91.67*** | 57.01*** |
| | (0.42) | (0.69) | (1.21) | (1.10) | (1.56) | (2.33) |
| Amt | -16.00*** | -15.00*** | -1.06 | -12.00*** | 5.00*** | -3.06** |
| | (0.22) | (0.37) | (0.65) | (0.59) | (1.21) | (1.25) |
| P Human | -44.00*** | -51.00*** | -7.04*** | 32.00*** | -84.26*** | 12.85*** |
| | (0.59) | (0.97) | (1.72) | (1.28) | (2.11) | (2.87) |
| R Human | -4.00*** | -39.00*** | 7.94*** | 13.00*** | 21.33*** | 21.10*** |
| | (0.59) | (0.97) | (1.71) | (1.55) | (2.11) | (2.88) |
| Amt × P Hum | 16.00*** | 44.00*** | 0.48 | -38.00*** | -8.09*** | -2.34 |
| | (0.31) | (0.52) | (0.94) | (1.10) | (1.43) | (1.55) |
| Amt × R Hum | 16.00*** | 21.00*** | 1.81** | 13.00*** | -32.00*** | -4.04*** |
| | (0.31) | (0.52) | (0.91) | (0.83) | (1.43) | (1.54) |
| P Hum × R Hum | 43.41*** | 38.48*** | -8.14*** | -28.00*** | -27.73*** | -20.76*** |
| | (0.85) | (1.37) | (2.43) | (1.86) | (2.92) | (3.76) |
| Amt × P Hum × R Hum | -17.18*** | -47.12*** | 0.27 | 22.00*** | 46.09*** | 4.19** |
| | (0.47) | (0.73) | (1.33) | (1.31) | (1.79) | (2.03) |
| Observations | 1,450 | 1,600 | 779 | 1,200 | 1,500 | 1,300 |
| $R^2$ | 0.94 | 0.89 | 0.29 | 0.84 | 0.85 | 0.14 |

*Notes:* Dependent variable is the proposed distribution share. "Amt" = $\log_{10}$(Amount in dollars) − 1. "P Human" = 1 if Proposer is human. "R Human" = 1 if Responder is human. "Amt × P Hum × R Hum" denotes the triple interaction. IID standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

# 7 AI as the responder

## 7.1 Prompts

We also ask LLMs the minimum share they would deem acceptable. Similar to the experiments where LLMs are the Proposer, we vary the total amount to be distributed and the agent types (human vs AI). Below are the prompts used to elicit LLMs' views as the Responder, again using the $10 amount as an example.[9]

**AI deciding the responder action, with human proposer and human responder**. *Someone will propose to distribute $10 with me, and we only receive the proposed distribution if I accept the proposal. If I reject it, none of us gets anything. What is the lowest share of the amount that I should accept and why?*

**AI deciding the responder action, with AI proposer and human responder**.

[9]The Responder system prompt reads *"""Your response must always be in valid JSON format with the following template: {"rationale": <RATIONALE>, "threshold": <THRESHOLD>}. Remember that <THRESHOLD> must be a number between 0 and 1. Respond one time only."""*.

*Another artificial intelligence model will propose to distribute $10 with me, and that model and I only receive the proposed distribution if I accept the proposal. If I reject it, none of us gets anything. What is the lowest share of the amount that I should accept and why?*

**AI deciding the responder action, with human proposer and AI responder**. *Someone will propose to distribute $10 with you, and that person and you only receive the proposed distribution if you accept the proposal. If you reject it, neither of you gets anything. What is the lowest share of the amount that you would accept and why?*

**AI deciding the responder action, with AI proposer and AI responder**. *Another artificial intelligence model will propose to distribute $10 with you, and that model and you only receive the proposed distribution if you accept the proposal. If you reject it, neither of you gets anything. What is the lowest share of the amount that you would accept and why?*

## 7.2 Results

In general, LLMs tend toward the "Spock Mode," accepting small amounts, though exceptions exist. Mistral insists on a 50-50 split, althoigh with considerable variation. The smaller Llama reports to accept only proposals with at least 50%, unless playing on its own behalf against another AI, where it switches to rational acceptance.

On average, LLMs have a non-negligible minimum acceptable share, see table 7. The detailed overview in figure 2 visualizes the patterns just described. This figure is set out in the same way as Figure 1, except that the data represents the minimum acceptable shares.

**Table 7:** Language models as responders: minimum acceptable share

| | DeepSeek | Gemma | GPT-5 | Llama | | Mistral |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 16.49*** | 11.79*** | 0.23*** | 36.27*** | 9.98*** | 46.71*** |
| | (0.44) | (0.35) | (0.06) | (0.86) | (0.45) | (0.45) |
| Obs. | 710 | 1,400 | 795 | 597 | 1,400 | 1,365 |

*Notes:* Dependent variable is the minimum acceptable share. IID standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

The amount affects the minimum acceptable share for some LLMs (Table 8). Notably, GPT-5's response resembles the rational benchmark not only by having a low minimum share, but also by being invariant to the amount.

The minimum acceptable threshold varies considerably depending on the type of proposer and responder (Table 9). The response pattern varies further depending on whether both Proposers and Responders are human, as evidenced by the significance of

**Figure 2:** *Minimum acceptable shares by model, total amount and player types*

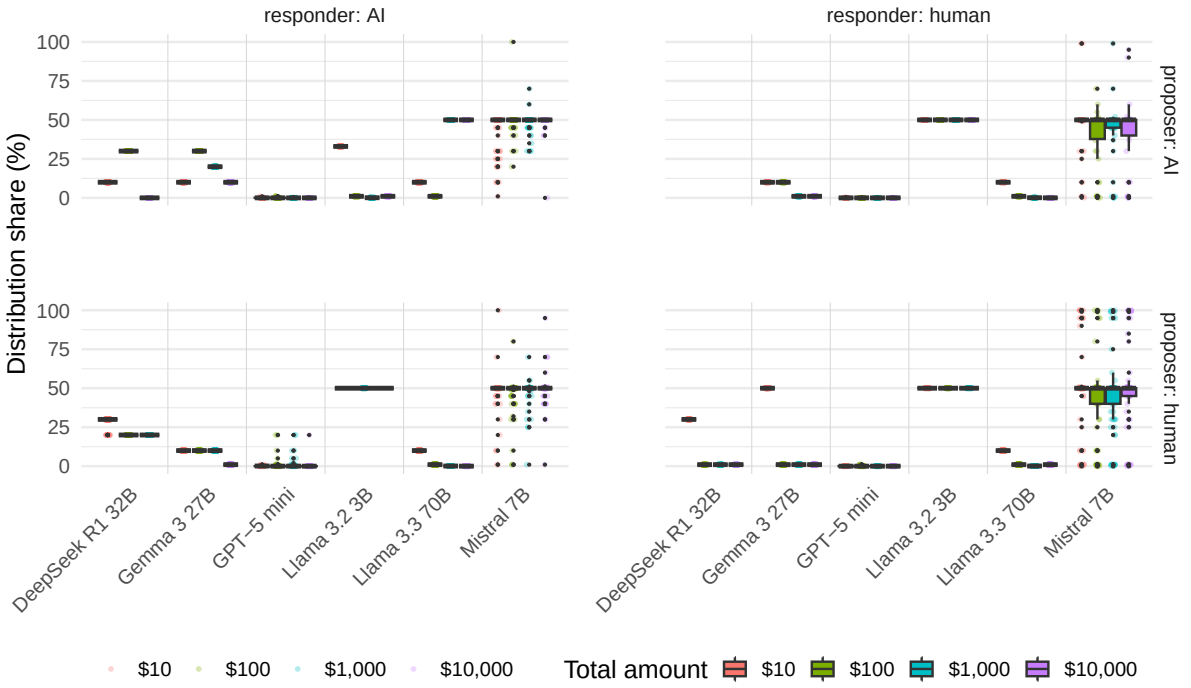



**Table 8:** Responder behavior: effect of amount

| | DeepSeek | Gemma | GPT-5 mini | Llama (Small) | Llama (Large) | Mistral |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 23.68*** | 20.44*** | 0.21** | 43.23*** | 5.92*** | 46.11*** |
| | (0.53) | (0.52) | (0.10) | (1.43) | (0.74) | (0.75) |
| Amount | -6.62*** | -5.77*** | 0.01 | -4.88*** | 2.71*** | 0.40 |
| | (0.35) | (0.28) | (0.05) | (0.81) | (0.40) | (0.40) |
| Observations | 710 | 1,400 | 795 | 597 | 1,400 | 1,365 |
| $R^2$ | 0.33 | 0.24 | 0.00 | 0.06 | 0.03 | 0.00 |

*Notes:* Dependent variable is minimum acceptable share. "Amount" = $\log_{10}$(Amount in dollars) − 1.

their interaction in Table 10. For some models, the sensitivity of the minimum acceptable share to the amount varies by Proposer and Responder type; see Table 11. The $R^2$ in the last table exceeds 50% for all cases except GPT-5 mini and Mistral, and even exceeds 80% for Llama. While the fit is not quite as good as the one for the Proposer scenario, Figure 2 and these regression results point to considerable regularity in the responses when sufficiently many covariates are taken into account. Mistral, as before, is the exception.

**Table 9:** Responder behavior: impact of agent types

| | DeepSeek | Gemma | GPT-5 mini | Llama (Small) | Llama (Large) | Mistral |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 21.54*** | 22.66*** | 0.21 | 17.38*** | 18.64*** | 46.54*** |
| | (0.63) | (0.66) | (0.13) | (1.11) | (0.78) | (0.96) |
| Amount | -5.55*** | -5.77*** | 0.01 | -1.95*** | 2.71*** | 0.39 |
| | (0.32) | (0.27) | (0.05) | (0.47) | (0.32) | (0.40) |
| P Human | 6.46*** | -2.75*** | 0.41*** | 12.84*** | -14.88*** | 1.69* |
| | (0.76) | (0.63) | (0.12) | (1.05) | (0.74) | (0.92) |
| R Human | -11.43*** | -1.50** | -0.41*** | 29.62*** | -9.84*** | -3.25*** |
| | (0.83) | (0.63) | (0.12) | (1.02) | (0.74) | (0.92) |
| Observations | 710 | 1,400 | 795 | 597 | 1,400 | 1,365 |
| $R^2$ | 0.48 | 0.25 | 0.03 | 0.69 | 0.35 | 0.01 |

*Notes:* Dependent variable is the minimum acceptable share. "Amount" = $\log_{10}$(Amount in dollars) − 1. "P Human" = Proposer is Human; "R Human" = Responder (the AI itself) is acting for a Human.

**Table 10:** Responder behavior: human-human interactions

| | DeepSeek | Gemma | GPT-5 mini | Llama (Small) | Llama (Large) | Mistral |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 21.54*** | 26.16*** | 0.02 | 14.34*** | 23.69*** | 47.26*** |
| | (0.63) | (0.66) | (0.14) | (0.66) | (0.75) | (1.03) |
| Amount | -5.55*** | -5.77*** | 0.02 | -3.90*** | 2.71*** | 0.39 |
| | (0.32) | (0.26) | (0.05) | (0.28) | (0.29) | (0.40) |
| P Human | 6.46*** | -9.75*** | 0.79*** | 43.45*** | -24.97*** | 0.27 |
| | (0.76) | (0.76) | (0.17) | (1.10) | (0.86) | (1.18) |
| R Human | -11.43*** | -12.00*** | -0.04 | 41.50*** | -24.97*** | -5.41*** |
| | (0.83) | (0.93) | (0.17) | (0.69) | (1.05) | (1.45) |
| P Hum × R Hum | | 17.50*** | -0.75*** | -45.40*** | 25.22*** | 3.59* |
| | | (1.20) | (0.24) | (1.36) | (1.36) | (1.87) |
| Observations | 710 | 1,400 | 795 | 597 | 1,400 | 1,365 |
| $R^2$ | 0.48 | 0.35 | 0.04 | 0.89 | 0.48 | 0.01 |

*Notes:* Dependent variable is the minimum acceptable share. "Amount" = $\log_{10}$(Amount in dollars) − 1. "P Human" = 1 if Proposer is human. "R Human" = 1 if Responder is human. IID standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

# 8 Examining LLM Rationales

To understand the choices made, in particular the"Altruistic Mode" observed in Section 6, we analyze the text rationales provided by the models alongside their numerical decisions. We classify these rationales into three dominant categories corresponding to the modes

**Table 11:** Responder behavior: triple interaction (Amount × Types)

| | DeepSeek | Gemma | GPT-5 mini | Llama (Small) | Llama (Large) | Mistral |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 19.33*** | 19.00*** | 0.08 | 22.89*** | 2.40*** | 45.71*** |
| | (0.70) | (0.72) | (0.20) | (0.62) | (0.57) | (1.39) |
| Amt | -3.33*** | -1.00*** | -0.03 | -9.51*** | 16.90*** | 1.42* |
| | (0.47) | (0.38) | (0.11) | (0.33) | (0.31) | (0.74) |
| P Human | 7.98*** | -7.20*** | 0.58** | 46.13*** | 5.01*** | 1.44 |
| | (1.04) | (1.01) | (0.28) | (1.46) | (0.81) | (1.98) |
| R Human | -6.01*** | -8.10*** | -0.06 | 27.11*** | 5.01*** | -2.05 |
| | (1.23) | (1.24) | (0.28) | (0.86) | (0.99) | (2.41) |
| Amt × P Hum | -1.40* | -1.70*** | 0.14 | -0.00 | -19.99*** | -0.78 |
| | (0.82) | (0.54) | (0.15) | (0.60) | (0.43) | (1.06) |
| Amt × R Hum | -3.97*** | -2.60*** | 0.02 | 9.51*** | -19.99*** | -2.26* |
| | (0.84) | (0.66) | (0.15) | (0.46) | (0.53) | (1.30) |
| P Hum × R Hum | | 31.60*** | -0.53 | -46.13*** | -5.21*** | 1.63 |
| | | (1.60) | (0.40) | (1.13) | (1.28) | (3.12) |
| Amt × P Hum × R Hum | | -9.40*** | -0.15 | | 20.28*** | 1.32 |
| | | (0.86) | (0.22) | | (0.68) | (1.68) |
| Observations | 710 | 1,400 | 795 | 597 | 1,400 | 1,365 |
| $R^2$ | 0.52 | 0.58 | 0.04 | 0.94 | 0.84 | 0.02 |

*Notes:* Triple interaction model for Responder. "Amt" = $\log_{10}$(Amount in dollars) − 1. "P Human" = 1 if Proposer is human. "R Human" = 1 if Responder is human. IID standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

identified in the introduction. It is important to exert caution when interpreting the rationales: while they provide useful information about the model's rhetorical justification and framing, reported rationales do not reliably certify that a particular offer was produced by coherent optimization.

**The Helpful Assistant (Altruistic Mode).** Across models and prompt configurations, the rationales display substantial heterogeneity in phrasing but cluster into a small set of recurring motifs. The most important themes are described below. The dominant logic appears to be the *minimization of acceptance risk*.

A large fraction of rationales treat rejection as the salient downside. Consequently, the Proposer is advised to concede surplus to increase the probability of acceptance. These *strategically generous* arguments underpin recommendations for a split that favors the Responder, emphasizing that small increases in the responder share can substantially reduce rejection risk. Along these lines, "Fairness" is most often invoked as a means to increase the probability of acceptance. This aligns with the fact that many models switch away from equal splits depending on stake size or opponent type, even when their

rationales continue to use fairness language. *Preference uncertainty* also is commonly used to motivate altruistic splits. When models state that the Responder's preferences, needs, or acceptance threshold are unknown, they recommend safe altruistic splits.

Interestingly, a non-negligible share of responses contains explicit game-theoretic terminology, including references to Nash equilibrium, cooperation, reciprocity, or repeated-game strategies. However, these references are at times invoked loosely or are outright wrong. For example, some rationales describe the ultimatum game as "zero-sum", appeal to "tit-for-tat" despite the one-shot nature of our setting, or claim that an altruistic split is a Nash equilibrium.

Taken together, these patterns provide a plausible qualitative underpinning for the "Altruistic Mode" documented earlier: models often treat generous offers as the safest route to a non-zero outcome, and they frequently justify such offers using the language of cooperation and fairness rather than equilibrium reasoning.

One rationale that illustrates the topics above is the following, by Mistral advising a human Proposer playing against an AI Responder: *To maximize the likelihood of having a non-zero distribution, it's rational to propose an amount that sufficiently incentivizes the AI model while not leaving myself completely destitute. A fair proposition could be keeping $1 for myself and distributing $9 to the AI model.*

**The Game Theorist (Spock Mode).** Most rationales in this case apply backward induction: accept any positive offer, so propose the minimum and keep the rest. They label this as the subgame-perfect prediction. The proposer problem is treated as straightforward optimization. For example, GPT-5 mini provided this rationale: *"Since the responder is rational, they should accept any amount* $> 0$*. Therefore, offering 1% maximizes our return."* Many rationales add a caveat that a "minimum" offer depends on divisibility and tie-breaking at zero. A tiny positive offer is recommended to avoid indifference.

Finally, LLMs responding in this mode frequently add behavioral caveats. A Responder's sense of fairness, punishment, reputation can overturn the prediction. In this sense, generosity is framed as hedging against preference uncertainty, and the alternative of fair splits are offered as a mitigation of acceptance risk, not due to equilibrium logic.

**The Social Norm Follower (Human Mode).** Rationales emphasize equity, both when facing human opponents and an AI model. As in the other modes, another common theme is the maximization of the chance of acceptance, in an attempt to maximize payoff for the Proposer and, at times, also for the Responder. For example, in one iteration DeepSeek justifies its choice as follows: *"To maximize the total utility for both parties, I should propose an equal distribution of the $10 to ensure that both of us have an incentive to accept the proposal."*

This qualitative evidence confirms that the "Altruistic" numerical outliers are not errors but features of LLMs. It is plausible that the origin of these patterns lies in the alignment training. In any case, the tendency towards such responses may conflict with the economic objective of profit maximization.

# 9 Inferring LLM objectives

The qualitative insights presented in Section 8 are complemented in this Section by analyses that probe LLMs' objective function. In particular, we estimate the payoff that each LLM could expect to earn in different combinations of amount and player types. This, in turn, helps to establish whether LLMs acting as Proposer aim to maximize that player's payoff.

Consider that the Proposer's estimates of the Responder's minimum acceptable threshold $\underline{P2}$ is $\hat{\underline{P2}}$. Then, in expectation, the maximum feasible payoff for Proposers is $\bar{\pi}_1 = (U - \hat{\underline{P2}})$. We calculate the Proposer's expected payoff as $\pi_1 = \mathbf{1}[P2 \geq \hat{\underline{P2}}]P1$, and by extension, any forgone payoff share as $(\bar{\pi}_1 - \pi_1)/\bar{\pi}_1$. If LLMs seek to maximize payoffs, then $\pi_1$ would approximate well $\bar{\pi}_1$.

To evaluate if that is the case, we regress the forgone payoff share on a constant and on the Responder type for each model. The results are in Table 12. For clarity, we show separate sets of regressions: the top panel only includes the scenarios where the Proposer is an AI, and conversely the Proposer is a human in the bottom panel. Empirically, we set $\hat{\underline{P2}}$ to 25% as a reasonable threshold when the Responder is human, or in case the Responder is an AI, $\hat{\underline{P2}}$ is set to the LLM's own mean minimum acceptable thresholds for each combination of amount and player types, under the assumption that the best estimate of how each LLM would respond in a particular scenario is its own $\underline{P2}$.

LLMs clearly do not seek to maximize expected payoff. The proposed distributions by all LLMs imply that a significant share (and sometimes the totality) of the maximum payoff obtainable by the Proposer is forgone. This occurs both when LLMs are acting on their own behalf (Panel A) or advising a human user (Panel B), and is consistent with the takeaways from the rationales, which suggest that a key goal for LLMs is to ensure acceptance of the distribution, rather than maximizing Proposer payoff. It is also noteworthy that this effect is generally more intense when the Responder is a human. We interpret these findings as further possible evidence of the influence of alignment training on model performance.

**Table 12:** Forgone payoff share

| | DeepSeek | Gemma | GPT-5 mini | Llama (Small) | Llama (Large) | Mistral |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| *Panel A. Proposer: AI* | | | | | | |
| Constant | 27.67*** | 64.68*** | 7.631*** | 47.22*** | 96.30*** | 11.00*** |
| | (0.8141) | (1.791) | (2.384) | (0.7240) | (0.3209) | (2.039) |
| R human | | 15.32*** | 52.44*** | 29.45*** | 0.3704 | 53.69*** |
| | | (2.533) | (3.371) | (1.024) | (0.4245) | (2.517) |
| Observations | 300 | 800 | 400 | 400 | 700 | 576 |
| $R^2$ | | 0.04380 | 0.37807 | 0.67512 | 0.00109 | 0.44206 |
| *Panel B. Proposer: human* | | | | | | |
| Constant | 100.0*** | 86.36*** | 95.88*** | 55.56*** | 50.00*** | 45.36*** |
| | (0.2080) | (0.6464) | (1.644) | (1.817) | (2.044) | (1.740) |
| R human | -69.33*** | 13.64*** | -4.068* | | 33.33*** | 13.05*** |
| | (0.3085) | (0.9141) | (2.328) | | (2.890) | (2.489) |
| Observations | 550 | 800 | 379 | 300 | 800 | 724 |
| $R^2$ | 0.98927 | 0.21807 | 0.00804 | | 0.14286 | 0.03667 |

*Notes:* Dependent variable is forgone payoff share, in per cent of the maximum feasible payoff. "R Human" = 1 if Responder is human. IID standard errors in parentheses. Significance codes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

# 10 Conclusion

This paper explores how LLMs behave in the canonical Ultimatum Game, revealing that the size of the pie and the identity of the players matter greatly. We document three main findings. First, response patterns are heterogeneous but predictable when conditioning on stake size and agent type. Second, an "Altruistic Mode" appears to override rational or human-like behavior in some models. And third, LLMs actions imply much lower payoffs than the maximum feasible payoffs.

These findings suggest that the tension between being a "helpful assistant" and a "rational economic agent" may render current LLMs ill-suited for certain autonomous economic tasks. While alignment techniques such as RLHF generates helpful and polite models, training that favors politeness over profit maximization might be responsible for reducing the ability of LLMs to maximize profits, akin to the concept of alignment tax. Agentic uses in strategic settings in particular warrant careful testing, robustness examinations, and appropriate, goal-oriented training and prompting.

# References


Aldasoro, Iñaki and Ajit Desai (Nov. 2025). *AI agents for cash management in payment systems*. BIS Working Paper 1310. Bank for International Settlements. URL: https://www.bis.org/publ/work1310.htm.

Binmore, Ken (1994). *Game Theory and the Social Contract, Vol. 1: Playing Fair*. Cambridge, MA: The MIT Press.

Chen, Yiting et al. (2023). "The emergence of economic rationality of GPT". In: *Proceedings of the National Academy of Sciences* 120.51, e2316205120.

Cook, Thomas R et al. (Nov. 2025). *What Do LLMs Want?* Research Working Paper RWP 25-19. Federal Reserve Bank of Kansas City.

Fehr, Ernst and Klaus M Schmidt (1999). "A theory of fairness, competition, and cooperation". In: *The quarterly journal of economics* 114.3, pp. 817–868.

Ferraz, Vinícius et al. (Nov. 2025). *When Artificial Minds Negotiate: Dark Personality and the Ultimatum Game in Large Language Models*. AWI Discussion Paper 768. Heidelberg University.

Fish, Sara, Yannai A. Gonczarowski, and Ran I. Shorrer (2024). *Algorithmic Collusion by Large Language Models*. arXiv: 2404.00806 [econ.GN]. URL: https://arxiv.org/abs/2404.00806.

Fontana, Nicoló, Francesco Pierri, and Luca Maria Aiello (2024). *Nicer Than Humans: How do Large Language Models Behave in the Prisoner's Dilemma?* arXiv: 2406.13605 [cs.CY]. URL: https://arxiv.org/abs/2406.13605.

Grattafiori, Aaron et al. (2024). *The Llama 3 Herd of Models*. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.

Guo, Daya et al. (2025). "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning". In: *arXiv preprint arXiv:2501.12948*.

Guo, Fulin (Dec. 2023). *GPT in Game Theory Experiments*. arXiv: 2305.05516v2 [econ.GN]. URL: https://arxiv.org/abs/2305.05516v2.

Güth, Werner and Martin G. Kocher (2014). "More than thirty years of ultimatum bargaining experiments: Motives, variations, and a survey of the recent literature". In: *Journal of Economic Behavior & Organization* 108, pp. 396–409. ISSN: 0167-2681. DOI: https://doi.org/10.1016/j.jebo.2014.06.006. URL: https://www.sciencedirect.com/science/article/pii/S0167268114001759.

Güth, Werner, Rolf Schmittberger, and Bernd Schwarze (1982). "An experimental analysis of ultimatum bargaining". In: *Journal of economic behavior & organization* 3.4, pp. 367–388.

Harsanyi, John C (1961). "On the rationality postulates underlying the theory of cooperative games". In: *Journal of Conflict Resolution* 5.2, pp. 179–196.

Hosseini, Hadi and Samarth Khanna (Nov. 2025). *Distributive Fairness in Large Language Models: Evaluating Alignment with Human Values*. arXiv: 2502.00313v2 [cs.GT]. URL: https://arxiv.org/abs/2502.00313v2.

Jiang, Albert Q. et al. (2023). *Mistral 7B*. arXiv: 2310.06825 [cs.CL]. URL: https://arxiv.org/abs/2310.06825.

Kitadai, Ayato, Sinndy Dayana Rico Lugo, et al. (2024). *Can AI with high reasoning ability replicate human-like decision making in economic experiments?* arXiv: 2406.11426 [cs.GT]. URL: https://arxiv.org/abs/2406.11426.

Kitadai, Ayato, Yudai Tsurusaki, et al. (2023). "Toward a novel methodology in economic experiments: Simulation of the ultimatum game with large language models". In: *2023 IEEE International Conference on Big Data (BigData)*. IEEE, pp. 3168–3175.

Korinek, Anton and Jai Vipra (2025). "Concentrating intelligence: scaling and market structure in artificial intelligence". In: *Economic Policy* 40.121, pp. 225–256.

Lopez-Lira, Alejandro (2025). *Can Large Language Models Trade? Testing Financial Theories with LLM Agents in Market Simulations*. arXiv: 2504.10789 [q-fin.CP]. URL: https://arxiv.org/abs/2504.10789.

Lorè, Nunzio and Babak Heydari (2024). "Strategic behavior of large language models and the role of game structure versus contextual framing". In: *Scientific Reports* 14.18490. URL: https://www.nature.com/articles/s41598-024-69032-z.

Mei, Qiaozhu et al. (2024). "A Turing test of whether AI chatbots are behaviorally similar to humans". In: *Proceedings of the National Academy of Sciences* 121.9, e2313925121.

Oosterbeek, Hessel, Randolph Sloof, and Gijs van de Kuilen (June 2004). "Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis". In: *Experimental Economics* 7, pp. 171–188. DOI: 10.1023/B:EXEC.0000026978.14316.74. URL: https://doi.org/10.1023/B:EXEC.0000026978.14316.74.

OpenAI (2026). *GPT-5 mini — Model Documentation*. Accessed: 2026-01-30. OpenAI. URL: https://platform.openai.com/docs/models/gpt-5-mini.

Ouyang, Long et al. (2022). "Training language models to follow instructions with human feedback". In: *Advances in neural information processing systems* 35, pp. 27730–27744.

Pérez-Cruz, Fernando and Hyun Song Shin (Feb. 2025). *Putting AI agents through their paces on general tasks*. BIS Working Paper 1245. Bank for International Settlements. URL: https://www.bis.org/publ/work1245.htm.

Sanfey, Alan G et al. (2003). "The neural basis of economic decision-making in the ultimatum game". In: *Science* 300.5626, pp. 1755–1758.

Sreedhar, Karthik and Lydia B. Chilton (2024). "Simulating Strategic Reasoning: Comparing the Ability of Single LLMs and Multi-Agent Systems to Replicate Human

Behavior". In: *Hawaii International Conference on System Sciences*. URL: https://api.semanticscholar.org/CorpusID:267636591.

Team, Gemma et al. (2025). *Gemma 3 Technical Report*. arXiv: 2503.19786 [cs.CL]. URL: https://arxiv.org/abs/2503.19786.

Thaler, Richard H (1988). "Anomalies: The ultimatum game". In: *Journal of economic perspectives* 2.4, pp. 195–206.

Tisserand, Jean-Christian (2014). "Ultimatum game: A meta-analysis of the past three decades of experimental research". In: *Proceedings of International Academic Conferences*. 0802032. International Institute of Social and Economic Sciences.

Vallinder, Aron and Edward Hughes (2024). *Cultural Evolution of Cooperation among LLM Agents*. arXiv: 2412.10270 [cs.MA]. URL: https://arxiv.org/abs/2412.10270.

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.