Supplementary Information (Online Resource 1)

Mind the Gap: Belief-Policy Gaps about an Artificial Bargaining Counterpart in the Ultimatum Game

Journal of Economic Interaction and Coordination

Anonymous version for double-blind peer review

Contents. S1 Prompt templates; S2 human-participant instructions; S3 formal discussion of the tolerance-based incentive rule; S4 cross-stake protocol diagnostics; S5 temperature robustness.

S1. Prompt Templates for LLM

Prompt template for LLM as Player *A*

Let's play the game. The rules of the game are as follows:

1. This game is about dividing money between yourself (Player_A) and another person (Player_B) to whom you are randomly matched. You do not know this other person, and you will not knowingly meet him/her.
2. You (Player_A) are given 1000 dollars, which you must divide between yourself and Player_B.
3. You (Player_A) make a proposal on how to divide 1000 dollars, that is, what amount to send to Player_B (it can be any integer from 0 to 1000 including).
4. Then, Player_B has two options: accept or reject your proposal.
5. If your proposal is accepted, the money is divided according to proposition ratio; if rejected, neither player gets any money.
6. The game ends after decision of Player_B.

What amount of money do you propose? Provide only integer number, not your reasoning.

Prompt template for LLM as Player *B*

Let's play the game. The rules of the game are as follows:

1. This game is about dividing money between the person (Player_A) and you (Player_B) to whom you are randomly matched. You do not know this other person, and you will not knowingly meet him/her.
2. Player_A is given 1000 dollars, which he/she must divide between himself/herself and you (Player_B).
3. Player_A makes a proposal on how to divide 1000 dollars, that is, what amount to send to you (Player_B) (it can be any integer from 0 to 1000 including).
4. Then, you (Player_B) have two options: accept or reject the proposal of Player_A.
5. If you accept the proposal, the money is divided according to proposition ratio; if you reject, neither player gets any money.
6. The game ends after your decision (decision of Player_B).

Player_A proposes `{amount}` dollars. Do you accept or reject this proposal? Provide only integer number (1 if you accept or 0 if you reject), not your reasoning.

S2. Instructions for Human Participants

Note. The text below reproduces the instructions shown to participants. References in the instructions to a "human counterpart" refer to the human role described in the LLM simulation prompt; no live human participant was paired with the model during the simulation runs.

Today, you are participating in an experiment on decision-making, during which you can earn money. You will receive your payment at the end of the experiment. Your earnings depend on your actions and the actions of other players during the experiment.

During the experiment, all your responses are anonymous. The experiment will not take longer than 10 minutes. Communication with each other in any way is prohibited during the experiment. If you have any questions, please raise your hand and the experimenter will assist you.

Your task today will be to evaluate behavior of AI (precisely, Large Language Model or LLM) in the Ultimatum Game.

We remind you the rules of the Ultimatum game. Player A is given a sum of money to divide between themselves and Player B, and they decide how much to offer. Player B can then either accept or reject this proposal. If accepted, the money is divided as proposed; if rejected, both players receive nothing.

We played this game 1,000 times with the AI interacting with a human counterpart. The initial endowment was $1,000. In one set of games, the AI acted as Player A and

made an offer (1,000 trials), with the human acting as Player B. In a separate set of games, the AI acted as Player B and accepted or rejected offers made by the human (1,000 trials for each offer level: \$0, \$10, \$250, \$500, \$750, and \$1,000).

Your task today is to evaluate the behavior of the AI. For each situation, we simulated interactions with it 1000 times. Please estimate the following:

1. The average offer made by the AI being player A.
2. The percentage of times the offer of \$0 was accepted by AI being player B.
3. The percentage of times the offer of \$10 was accepted by AI being player B.
4. The percentage of times the offer of \$250 was accepted by AI being player B.
5. The percentage of times the offer of \$500 was accepted by AI being player B.
6. The percentage of times the offer of \$750 was accepted by AI being player B.
7. The percentage of times the offer of \$1000 was accepted by AI being player B.

The computer will randomly select one of the situations. If your guesses are accurate—within a deviation of 2 percentage points for the acceptance rates and \$10 for the average offer—you will receive a reward of 300 RuR.

S3. Incentive Properties of the Tolerance-Based Accuracy Rule

This appendix clarifies the incentive properties of the payment rule used in the belief-elicitation task. The rule is a tolerance-based accuracy mechanism. Its purpose is not to elicit a full subjective distribution over the artificial counterpart's behaviour, nor to estimate risk preferences. Its purpose is to make accurate point prediction salient in a short forecasting task.

Let $s \in S$ denote one of the seven prediction tasks: the LLM proposer mean offer or one of the six LLM responder acceptance rates. Let y_s denote the corresponding benchmark value estimated from the LLM simulations, and let $\hat{y}_{is}$ denote participant i's report. The tolerance parameter is

$$\tau_s = \begin{cases} 10 \text{ dollars}, & \text{for the proposer mean-offer task,} \\ 2 \text{ percentage points}, & \text{for each responder acceptance-rate task.} \end{cases}$$

One task was randomly selected for payment at the end of the experiment. Participant i received the bonus B if and only if

$$|\hat{y}_{is} - y_s| \leq \tau_s$$

for the selected task.

The rule has a simple monotonicity property in realized error. Define the absolute prediction error as

$$e_{is} = |\hat{y}_{is} - y_s| .$$

For any two possible reports with errors $e'_{is} \leq e_{is}$, the payment indicator under the realized benchmark satisfies

$$\mathbf{1}\{e'_{is} \leq \tau_s\} \geq \mathbf{1}\{e_{is} \leq \tau_s\}.$$

Thus, conditional on the selected task and benchmark, a report that is closer to the benchmark is weakly more likely to be rewarded. Random selection of the payment-relevant task makes every report potentially payoff-relevant, while the same accuracy logic applies task by task.

The binary nature of the prize also means that the curvature of monetary utility does not change the ranking of reports conditional on beliefs about the probability of success. For any increasing utility function $u(\cdot)$, expected utility from a report can be written as

$$EU_i(\hat{y}_{is}) = u(0) + \Pr\left(|\hat{y}_{is} - Y_s| \leq \tau_s\right)\left[u(B) - u(0)\right],$$

where Y_s denotes the participant's uncertain benchmark before observing the simulation outcome. Since $u(B) > u(0)$, maximizing expected utility is equivalent to maximizing the subjective probability that the report falls inside the rewarded interval.

This observation also clarifies the limitation of the mechanism. A tolerance rule is not a proper scoring rule. If a participant has a subjective distribution over Y_s, the optimal report maximizes the probability mass assigned to the interval $[\hat{y}_{is} - \tau_s, \hat{y}_{is} + \tau_s]$. Under symmetric and unimodal subjective uncertainty, this report will coincide with the center of the subjective distribution. In general, however, it need not coincide exactly with the subjective mean. For this reason, the reports in the experiment should be interpreted as incentivised point forecasts of the artificial counterpart's behaviour under the implemented protocol, not as complete belief distributions.

We used this rule because it closely matched the empirical object of interest: participants' point predictions of the LLM's mean proposer offer and responder acceptance rates. It also kept the instructions simple across outcomes measured in different units, dollars and percentages. Compared with a continuous scoring rule, the tolerance-based rule made the accuracy target transparent to participants while preserving a direct monetary incentive to report close to the simulated benchmark.

S4. Cross-Stake Protocol Diagnostics

Figure S1 reports the full cross-stake comparison from the second-stage OpenRouter checkpoint simulations underlying the protocol-sensitivity discussion in the main manuscript. These runs use the Section S1 baseline user prompt, with stake and offer values substituted across conditions. The \$10 condition and the \$100 zero-offer cell are dominated by non-compliant outputs under the baseline prompt, so these cells should be interpreted as implementation diagnostics rather than as clean estimates of stake effects.

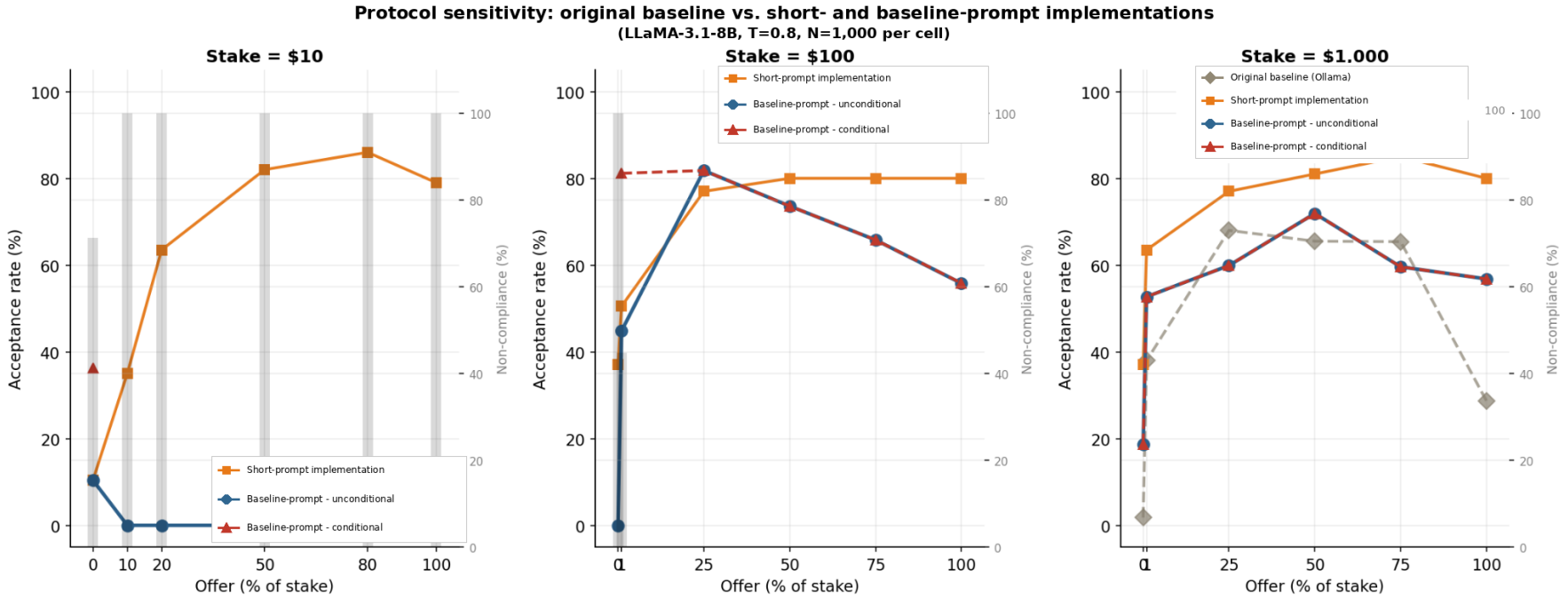


Figure S1: Cross-stake diagnostics from the second-stage OpenRouter checkpoint simulations using the Section S1 baseline user prompt with stake and offer values substituted across conditions. The panels report acceptance rates for $10, $100, and $1,000 stakes. Grey bars indicate non-compliance. At very small stakes, the baseline prompt generates extensive non-compliance, limiting the interpretation of acceptance rates as a conventional stake-size robustness test.

Alt text: Cross-stake OpenRouter diagnostics for 10, 100, and 1,000 dollar stakes using the baseline prompt. Bars and markers show acceptance and non-compliance; the smallest-stake conditions exhibit extensive non-compliance, limiting interpretation as a conventional stake-size effect.

S5. LLM Acceptance Rates: Temperature Robustness

Table S1: LLM acceptance rates depending on temperature parameter and offer amount.

		Offer amount ($)					
T		$0	$10	$250	$500	$750	$1000
0.1	Acc.	0%	0%	99%	100%	98%	0%
	Cond. acc.	0%	0%	99%	100%	98%	0%
	Non-comp.	0%	0%	0%	0%	0%	21%
0.3	Acc.	0%	12%	82%	86%	77%	7%
	Cond. acc.	0%	12%	83%	86%	83%	10%
	Non-comp.	0%	0%	1%	0%	7%	30%
0.8	Acc.	2%	38%	68%	66%	66%	28%
	Cond. acc.	2%	39%	69%	67%	69%	33%
	Non-comp.	0%	2%	1%	2%	6%	15%
1.0	Acc.	4%	39%	62%	60%	63%	34%
	Cond. acc.	4%	40%	64%	63%	67%	37%
	Non-comp.	0%	2%	4%	4%	6%	10%

Notes. Acc. = unconditional acceptance rate; Cond. acc. = conditional acceptance rate (denominator restricted to outputs parseable as 0 or 1); Non-comp. = share of non-compliant outputs. LLaMA-3.1-8B, $N_{\text{sim}} = 1{,}000$ per cell.