Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Tatyana Zhuravleva
HSE University, St. Petersburg
Amina Marnova
HSE University, St. Petersburg
Andrea Morone
andrea.morone@uniba.it

University of Bari Aldo Moro
Andrei Ternikov
HSE University, St. Petersburg

Research Article

Keywords:

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11409231/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

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

Abstract

This study examines how people form beliefs about an artificial bargaining counterpart and whether misperceiving its observed policy has strategic costs. We elicit incentivised beliefs from 108 university students about a large language model's proposer offers and responder acceptance decisions in an Ultimatum Game and benchmark them against 1,000 simulations per scenario. The average proposer forecast shows no detectable mean bias, but individual point forecasts are heterogeneous. Responder beliefs diverge systematically from the observed acceptance policy: 85.2% of participants report a weakly monotone acceptance schedule even though the baseline artificial policy is non-monotone. These gaps affect belief-implied best responses: \$500 belongs to the maximizing set for 62.0% of participants, whereas the baseline policy makes \$250 optimal. Counterfactual best responding to elicited beliefs therefore generates sizeable expected payoff losses. Additional simulations show that the observed artificial policy is sensitive to prompt formulation, decoding temperature, numerical scale, and inference environment. The results connect experimental bargaining with the economics of interacting heterogeneous agents by showing how mismatches between human beliefs and protocol-contingent artificial policies can create payoff-relevant frictions in human–AI interaction.

Keywords: Ultimatum Game, Belief-Policy Gaps, Artificial Agents, Experimental Economics, Large Language Models, Human–AI Interaction.
JEL Classification: C70, C91, D91, O33

1 Introduction

Large language models (LLMs) are increasingly embedded in everyday applications and are now routinely encountered as conversational partners, advisors, and decision aids. In parallel, a growing methodological debate in the social sciences concerns whether LLMs can serve as stand-ins for human participants in behavioural and game-theoretic experiments, given their scalability, speed, and low marginal cost (e.g., Park et al., 2022; Argyle et al., 2023; Aher et al., 2023). At the same time, accumulating evidence documents substantial cross-model variability and systematic departures from human behaviour, leaving the scope conditions for using LLMs in strategic settings unsettled. Notably, Fan et al. (2024) show that LLMs may struggle to instantiate preferences under uncommon assumptions, update beliefs even from simple information, and act consistently on those beliefs. These findings counsel caution in substituting LLMs for humans.

A distinct but closely related issue arises when humans interact with artificial agents as strategic counterparts. In such environments, the relevant object is not only how

an LLM behaves, but also how human players believe it behaves. Strategic outcomes depend on beliefs: miscalibrated expectations can distort best responses, shift surplus, and generate payoff losses. As artificial agents increasingly enter bargaining, negotiation, and decision-support environments, human players face a new form of strategic uncertainty. They must form beliefs not about another human player's preferences, type, or intentions, but about the observable behavioural policy implemented by an algorithmic system under a given interaction protocol.

This paper studies that problem. We ask how participants forecast the observed policy of an artificial bargaining counterpart under a fixed Ultimatum Game protocol, and whether gaps between perceived and observed policies are payoff-relevant. We treat this as a question about beliefs over an observable artificial policy, not as a claim that belief errors are uniquely AI-specific. LLM behaviour can be highly protocol-sensitive: it may depend on prompting, identity disclosure, instruction-following constraints, refusal or guardrail mechanisms, and numerical or contextual features of the prompt. As a result, the mapping from a game description to observed play may differ from the mapping generated by human preferences, norms, or strategic reasoning. The key question is therefore not simply whether people are inaccurate, but which behavioural policy they think they face, and whether misperceiving that policy changes optimal strategic behaviour.

Misperceptions about strategic counterparts are not new: the human-human bargaining literature shows that beliefs about others' behaviour can be imperfect and payoff-relevant. What may be distinctive about artificial counterparts is that the mapping from strategic environment to observed action can be shaped by protocol features that differ from those governing human behaviour: prompt wording, identity disclosure, instruction-following constraints, refusal or non-compliance patterns, and numerical representation of the task. Thus, the relevant question is not whether humans ever misforecast strategic counterparts, but whether they project a familiar human-human bargaining model onto an artificial agent whose observed policy is protocol-contingent. We therefore ask whether and where participants' beliefs about an LLM responder depart from its observed policy, and whether these belief-policy gaps distort best responses and generate payoff losses.

We study this issue in the ultimatum game (UG), which is particularly well suited to identifying the economic consequences of belief-policy gaps. The UG provides a simple mapping from beliefs about an artificial counterpart's behaviour into optimal strategic choices: when the artificial agent acts as responder, its acceptance policy $q(x)$ directly determines the expected payoff from each possible offer, so that misperceiving this policy can change the proposer's best response and generate a measurable payoff loss. The game is also role-separable, allowing beliefs about the model's proposer behaviour and responder policy to be elicited independently. Finally, the UG provides an unusually well-established human-human benchmark: human responders typically display a broadly monotone acceptance pattern, with low offers rejected more frequently and fair or generous offers accepted more often. This makes the setting especially informative for studying whether people project a familiar human bargaining heuristic onto an artificial counterpart whose observed policy need not follow the same pattern. Crucially, none of these advantages requires attributing a utility function or monetary motivation to the LLM: the relevant object is the observable policy generated under a fixed protocol and the beliefs humans form about that policy.

To anchor this comparison, it is useful to begin from the human-human benchmark. The Ultimatum Game is especially useful for our purposes because human-human behaviour in this environment is unusually well documented. The standard game-theoretic prediction

under selfish monetary preferences is simple: proposers should offer the smallest positive amount and responders should accept any positive offer. Decades of laboratory evidence, however, show systematic departures from this benchmark. Human proposers commonly offer substantial shares of the pie, often close to an equal split, and responders frequently reject low or unfair offers. In their meta-analysis of 37 studies and 75 results, Oosterbeek et al. (2004) report that proposers offer on average about 40 percent of the pie and that rejection rates decline as the share offered to the responder increases. Thus, in human-human ultimatum bargaining, the empirical acceptance pattern is typically increasing in the offer share: low offers are more likely to be rejected, while fair or generous offers are usually accepted.

This evidence gives us an external benchmark for interpreting beliefs about an artificial responder. The benchmark is not meant to replace a matched human-responder treatment in our own experiment. Rather, it identifies a well-established human-human regularity that participants may plausibly use as a mental model when they reason about an unfamiliar strategic counterpart. A natural human-like heuristic is monotone: low offers should be rejected relatively often, intermediate or fair offers should be accepted more often, and very generous offers should be accepted almost surely. Such a heuristic is sensible in human-human Ultimatum Games, where rejection behaviour is commonly interpreted as punishment for unfairness, social-preference expression, or norm enforcement.

The same human-human literature also shows why the Dictator Game is an informative background comparison. Dictator Game experiments demonstrate that allocations in simple distributional environments are not purely selfish (Engel, 2011). At the same time, comparisons between Dictator and Ultimatum Games show that bargaining behaviour cannot be reduced to pure distributional preferences: strategic concerns and anticipated rejection risk also matter when responders can reject offers (Forsythe et al., 1994). This distinction is important for our design because beliefs about the responder's acceptance policy directly discipline the proposer's optimal offer.

A more closely related strand of human-human work studies beliefs about other human players. Several experiments elicit expectations about human counterparts' offers, rejection thresholds, or minimum acceptable offers. Azar et al. (2015), for example, elicit proposers' beliefs about the minimum amount responders will accept and responders' beliefs about both expected offers and other responders' acceptance thresholds in a multi-period Ultimatum Game. They show that beliefs are not perfectly calibrated and may remain sticky even after feedback. Trautmann and van de Kuilen (2015) compare belief-elicitation mechanisms in a simple two-player strategic setting and show that elicitation procedures can affect the predictive content of reported beliefs. Vavra et al. (2018) further show that expectations about upcoming offers affect responders' acceptance decisions: the mean expected offer shifts the acceptance threshold, while the expected variance affects how strictly responders apply that threshold. This evidence makes clear that predicting human strategic behaviour is itself a non-trivial task.

Our study extends this belief-elicitation perspective from human-human bargaining to artificial strategic counterparts. Existing human-human evidence provides a contextual benchmark, not an internal control condition. We therefore do not claim to identify AI-specific miscalibration in the strongest sense. Instead, we ask whether participants' beliefs about an artificial responder resemble the familiar monotone acceptance pattern from human-human bargaining, and whether departures between those beliefs and the observed LLM policy distort best responses.

A growing empirical literature examines AI behaviour in economic environments and

how humans perceive, trust, and strategically respond to algorithmic agents. Some studies assess whether LLMs satisfy standard rationality criteria in choice or game play (e.g., Brookins and DeBacker, 2024). More directly, Barak and Costa-Gomes (2025) provide controlled monetarily-incentivised laboratory evidence on human behaviour against LLM opponents in a multi-player p-beauty contest, showing systematic shifts in choices when subjects face LLMs rather than humans and linking these shifts to stated beliefs about LLM reasoning. Complementary evidence highlights that attitudes toward AI and behavioural responses may diverge and may depend on disclosure and context. For instance, Ozkes et al. (2024) show that people report preferring human over algorithmic partners, while their revealed strategic behaviour is often similar across counterpart types except in specific beneficiary structures; and Buchanan and Hickman (2024) show that explicit AI attribution can reduce trust and increase fact-checking. Taken together, existing work clarifies AI behavioural patterns and contextual determinants of trust, but systematic evidence on what people believe a specific LLM will do in a canonical bargaining game, and on the payoff consequences of those beliefs, remains limited.

Our study addresses this gap by measuring incentivised beliefs about an unspecified artificial counterpart's play in the ultimatum game and benchmarking those beliefs against the observed behaviour of LLaMA-3.1-8B under a fixed prompt and decoding protocol. The design proceeds in two steps. First, we elicit the LLM's behaviour as proposer and responder using a fixed instructed prompt and 1,000 independent simulations per scenario, yielding stable estimates of the model's offer distribution and acceptance frequencies. Second, we elicit human expectations about how the model behaves in the same scenarios. University students ($N = 108$) predict (i) the model's mean offer as proposer and (ii) the model's acceptance rates as responder for six fixed offers (0, 10, 250, 500, 750, 1000). Accuracy is incentivised using a tolerance-based reward scheme described in Section 3. Participants were informed that the counterpart was an AI/LLM but were not told the exact model identity; hence, the elicited reports should be interpreted as beliefs about a generic artificial counterpart, benchmarked ex post against a specific LLaMA-3.1-8B policy.

A key advantage of our approach—relative to designs where humans directly play against an LLM—is that it isolates beliefs about the model's behaviour from confounds that otherwise complicate interpretation. First, participants typically understand that an LLM does not meaningfully value monetary payoffs, whereas monetary incentives are the canonical motivational foundation of laboratory games. Second, direct play against an LLM collapses two distinct drivers of observed choices into a single behavioural object: the player's own distributive preferences and the player's expectations about the LLM's response. By shifting the object of measurement from strategic play with the model to incentivised belief elicitation, we obtain a cleaner measure of perceived policy that is not mechanically confounded with participants' own social preferences and does not rely on the premise that the model is money-motivated.

To discipline interpretation of the beliefs data, we introduce an explicit beliefs-as-policy framework in Section 2. We treat the LLM as implementing observable proposer and responder policies that can depend on an interaction protocol θ, capturing identity disclosure and other informational features of the prompt. Human beliefs are represented by perceived policy objects. This framework clarifies what the beliefs data identify: not the LLM's internal "utility function", training corpus, or guardrails separately, but the behavioural policy subjects expect to face. It also allows us to connect belief-policy gaps to strategic consequences by comparing the best response to the observed LLM acceptance

policy with best responses implied by perceived acceptance policies, distinguishing the aggregate policy from individual-level best responses.

Empirically, we document a systematic gap between perceived and observed LLM behaviour in bargaining, but the nature of the gap differs sharply by role. For the proposer role, the average participant forecast is \$465, compared with a simulated mean offer of \$476; a two-sided one-sample test does not detect an aggregate mean bias ($t(107) = -0.43$, $p = 0.669$, 95% CI for the mean forecast: [\$413.9, \$516.1]). The relevant proposer-side pattern is therefore cross-subject heterogeneity in point forecasts, not evidence that participants perceive the model's offer distribution as more dispersed (Figure 1). For the responder role, by contrast, miscalibration is substantial in levels and varies across the offer grid: participants underestimate acceptance for some low and intermediate offers and dramatically overestimate acceptance at the highest offer (Table 1 and Figure 3). At the same time, the beliefs are highly structured: 92 of 108 participants (85.2%) report a weakly non-decreasing perceived acceptance schedule, and beliefs at adjacent offer levels are strongly correlated. This pattern is consistent with participants imposing a familiar human-like monotone acceptance logic on an artificial responder whose baseline observed policy is itself non-monotone.

Crucially, these belief-policy gaps are strategically consequential at both the aggregate and individual levels. On the elicited offer grid, the baseline observed LLM acceptance policy implies a payoff-maximising offer of \$250, whereas the average perceived acceptance policy implies an equal split of \$500. The individual-level analysis shows that this result is not an artefact of averaging heterogeneous beliefs: among the 101 participants with a unique belief-implied best response, \$500 is optimal for 60 (59.4%), and it belongs to the maximizing set of 67 of all 108 participants (62.0%). Among participants with a unique best response, the mean payoff loss under the baseline observed policy is \$168.4 and the median is \$181.8. Thus, the belief-policy gap is not only a matter of perception or trust; for many participants it changes the action implied by their own beliefs and, under counterfactual best responding, would generate economically meaningful payoff losses.

Our contribution is threefold. First, we extend the belief-elicitation approach used in human-human bargaining to an artificial strategic counterpart, measuring incentivised point forecasts about an LLM's proposer and responder policies. Second, we use the extensive human-human Ultimatum Game literature as a contextual benchmark and show at the individual level that most participants impose a monotone acceptance schedule on the artificial responder. Third, we quantify the counterfactual payoff consequences of belief-policy gaps at both aggregate and individual levels, showing that the equal-split best response is widespread across participants and that belief-implied best responses would generate sizeable payoff losses under the baseline observed policy. More broadly, the paper connects experimental bargaining with the study of interacting heterogeneous agents: strategic outcomes depend on the interaction between a protocol-contingent artificial policy and heterogeneous human beliefs about that policy.

The paper proceeds as follows. Section 2 presents the beliefs-as-policy framework and derives testable predictions. Section 3 describes the experimental design, the simulation protocol, and the incentivised belief elicitation. Section 4 reports the main results on aggregate and individual-level miscalibration and quantifies the best-response distortion and payoff loss implied by belief errors. Section 5 examines the sensitivity of LLM responder behaviour to prompt formulation, numerical scale, and inference environment. Section 6 discusses implications. Section 7 concludes.

2 A Beliefs-as-Policy Framework

This section introduces a parsimonious framework to interpret the beliefs data and to connect (mis)calibration to payoff-relevant strategic implications. The key modelling choice is deliberately reduced-form: we treat the LLM as implementing observable behavioural policies—a proposer offer-generation policy and a responder acceptance policy—rather than committing to any specific micro-foundation (utility, training data, equilibrium knowledge, or guardrails). Importantly, the experimental prompts do not assign the LLM an explicit objective function, such as maximising its own monetary payoff, total surplus, or distributive fairness. This is a feature of the design: the policy objects introduced below capture the model's observed prompt-conditioned behaviour without imposing an economic utility function on an artificial agent whose monetary incentives are not naturally defined.

2.1 Protocols and reduced-form policy objects

A central primitive is the interaction protocol $\theta \in \Theta$, capturing the information structure induced by prompts and instructions (e.g., whether identity is disclosed: LLM vs. LLM, human vs. LLM, or unspecified). Allowing for θ makes portability explicit: the same LLM can implement different observable behaviours under different protocols.

LLM as proposer. Let $X \subseteq [0,1]$ denote the finite offer grid (as implemented in the experiment).

Definition 1 (True offer policy and its moments)**.** *For each protocol θ, define the LLM proposer's offer-generation policy as*

$$g_\theta(x) \equiv \Pr\left(\textit{LLM offers } x \mid \theta\right), \qquad x \in X. \tag{1}$$

Let $\mu_\theta \equiv \mathbb{E}_{g_\theta}[x]$ and $\sigma^2_\theta \equiv \mathrm{Var}_{g_\theta}(x)$ denote the mean and variance of offers implied by $g_\theta(\cdot)$. In the data, $g_\theta(\cdot)$ is estimated from repeated independent LLM simulations under protocol θ.

Definition 2 (Beliefs about the proposer mean)**.** *Let $\hat{\mu}_{i,\theta}$ denote subject i's incentivised prediction of μ_θ. Define the average belief and cross-subject dispersion as*

$$\hat{\mu}_\theta \equiv \mathbb{E}[\hat{\mu}_{i,\theta}], \qquad \hat{v}_\theta \equiv \mathrm{Var}(\hat{\mu}_{i,\theta}). \tag{2}$$

Even if $\hat{\mu}_\theta$ is close to μ_θ, the cross-subject distribution of reports remains informative. A concentrated true policy (small σ^2_θ) together with a dispersed cross-subject distribution (large $\hat{v}_\theta$) indicates heterogeneity across subjects in point beliefs. It should not be interpreted as within-subject uncertainty or perceived noisiness unless subjective distributions are directly elicited.

LLM as responder.

Definition 3 (True acceptance policy)**.** *For each protocol θ, define the LLM responder's acceptance policy as*

$$q_\theta(x) \equiv \Pr\left(\textit{LLM accepts offer } x \mid \theta\right), \qquad x \in X. \tag{3}$$

Definition 4 (Belief acceptance policy)**.** *Let $\hat{q}_{i,\theta}(x)$ denote subject i's incentivised belief about acceptance:*

$$\hat{q}_{i,\theta}(x) \equiv \Pr_i\left(\textit{LLM accepts offer } x \mid \theta\right), \qquad \hat{q}_\theta(x) \equiv \mathbb{E}\left[\hat{q}_{i,\theta}(x)\right]. \tag{4}$$

2.2 Responder-side best responses and payoff consequences

Consider the ultimatum game with a unit pie. The optimisation problem considered here belongs to the human proposer, not to the artificial responder: the LLM is not instructed to maximise a monetary objective, and $q_\theta(x)$ is the empirically observed acceptance policy generated under protocol θ. If a human proposer chooses offer x against that policy, the proposer's true expected payoff is

$$\pi_\theta(x) = (1-x)q_\theta(x). \tag{5}$$

The payoff-maximising offer is

$$x^\star_\theta \in \arg\max_{x\in X}(1-x)q_\theta(x). \tag{6}$$

A subject who best-responds to her beliefs evaluates instead $\hat{\pi}_{i,\theta}(x) = (1-x)\hat{q}_{i,\theta}(x)$ and chooses $\hat{x}_{i,\theta} \in \arg\max_{x\in X}(1-x)\hat{q}_{i,\theta}(x)$.

Proposition 1 (Belief-based optimisation and payoff loss)**.** *Fix θ. Define the loss from belief-based play as*

$$\Delta_{i,\theta} \equiv \pi_\theta(x^\star_\theta) - \pi_\theta(\hat{x}_{i,\theta}) = (1-x^\star_\theta)q_\theta(x^\star_\theta) - (1-\hat{x}_{i,\theta})q_\theta(\hat{x}_{i,\theta}). \tag{7}$$

Then $\Delta_{i,\theta} \geq 0$ for all i. Moreover, the wedge between perceived and true expected payoffs is

$$\hat{\pi}_{i,\theta}(x) - \pi_\theta(x) = (1-x)\left(\hat{q}_{i,\theta}(x) - q_\theta(x)\right). \tag{8}$$

The payoff analysis below does not require a general sufficient condition for over-offering. It directly computes best responses on the elicited offer grid using the observed LLM acceptance policy and the perceived policies elicited from participants.

2.3 Predictions and empirical mapping

The framework yields three main testable predictions.

Prediction 1 (Proposer: aggregate mean bias and cross-subject heterogeneity). (i) Test whether the average point forecast $\hat{\mu}_\theta$ differs from the simulated mean μ_θ. (ii) Characterize heterogeneity in the cross-subject distribution of point forecasts $\hat{\mu}_{i,\theta}$. Because participants report a point forecast of the mean rather than a subjective distribution of offers, cross-subject dispersion should not be interpreted as perceived within-model variability.

Prediction 2 (Responder: policy miscalibration across offers). Average beliefs $\hat{q}_\theta(x)$ differ systematically from $q_\theta(x)$, with pronounced miscalibration at extremes.

Prediction 3 (Best-response distortion and payoff loss). When $\hat{q}_{i,\theta} \neq q_\theta$, belief-based optimisation can imply $\hat{x}_{i,\theta} \neq x^\star_\theta$ and a non-trivial payoff loss $\Delta_{i,\theta}$ (Proposition 1). We examine protocol dependence descriptively in Section 5, where we report how the observed LLM responder policy varies with prompt formulation, numerical scale, and inference environment.

3 Experimental Design and Procedure

We investigated the behavior of large language models (LLMs) in the Ultimatum Game, a standard paradigm in behavioral economics.

The study consisted of two stages: (i) LLM simulations in which the model acted either as proposer or responder in a prompt describing a human counterpart; and (ii) incentivised human predictions of the LLM's behaviour.

In the proposer role, the model determined an integer offer between \$0 and \$1000 to a human counterpart described in the prompt. In the responder role, the model decided whether to accept or reject fixed offers of \$0, \$10, \$250, \$500, \$750, and \$1000 made by a human proposer described in the prompt. No live human participant was paired with the model during these simulation runs. Each of the seven scenarios (one proposer, six responder) was simulated independently 1,000 times to obtain stable estimates of behavior.

Simulation prompts were adapted from the baseline strategy of Brookins and DeBacker (2024), applied as-is without additional alignment or fine-tuning, consistent with findings in Fan et al. (2024). The simulation script generated observations at four temperature settings, $T \in \{0.1, 0.3, 0.8, 1.0\}$. The main human-belief analyses use the $T = 0.8$ simulations as the baseline benchmark, following Lore and Heydari (2024); the remaining temperature conditions are used for the sensitivity analysis reported in Online Resource 1, Section S5.

The prompts deliberately specify the strategic environment and its monetary consequences without assigning the model an explicit objective such as maximising its own payoff, maximising joint surplus, or implementing fairness. As Online Resource 1, Section S1 shows, the proposer is simply asked what amount it proposes, while the responder is asked whether it accepts or rejects a given offer. This design allows us to measure the behavioural policy generated by the model under a fixed interaction protocol, rather than confounding that policy with an objective function imposed by the experimenter.

We employed the open-weight LLaMA-3.1-8B model (accessed via the Ollama interface) because it provides a controllable benchmark for repeated simulations under a specified prompting and decoding protocol. More specifically, the baseline simulations used the model tag `llama3.1:8b`, Ollama model digest `46e0c10c039e`, and Ollama v0.14.0, and were run on 21 January 2026. No system prompt was used, so the model received no additional role, objective, or behavioural instruction beyond the user-level Ultimatum Game prompts reproduced in Online Resource 1, Section S1. No random seed was explicitly set. Each generation started from a fresh context consisting of a newly constructed one-message input, with no prior conversation history reused across simulations. The script explicitly set temperature and `stream=False`; all remaining generation parameters were left at their Ollama default settings.

Participants were not told that the benchmark model was LLaMA-3.1-8B. The belief-elicitation task therefore measures beliefs about an AI/LLM counterpart described generically in the instructions, which we benchmark against the behaviour of this specific model under the implemented protocol. This feature mirrors many real-world interactions with AI systems, where users know they are interacting with an artificial system but not the exact model version.

Full prompt texts are reported in Online Resource 1, Section S1.

In the second stage, 108 human subjects, third-year Economics students at a large university in St. Petersburg, Russia, were asked to predict the LLM's behavior. The

task was implemented in oTree, and the human-subject data were collected on May 22 and May 26, 2025. For the proposer scenario, participants estimated the model's mean offer, with an allowable error margin of $10. For each responder scenario, they predicted the proportion of acceptance decisions, with an error margin of 2 percentage points. To incentivize accuracy, one scenario per subject was randomly selected ex post, and a prize of 300 rubles was awarded to participants whose predictions fell within the margin. The instructions provided to participants are reproduced in Online Resource 1, Section S2.

3.1 Incentive mechanism

The belief-elicitation task was incentivised using a tolerance-based accuracy rule. The mechanism was designed for simple point forecasts of simulated behavioural quantities, not for eliciting complete subjective distributions. At the end of the experiment, one of the seven prediction tasks was randomly selected for payment. The participant received the bonus if the selected report was sufficiently close to the corresponding benchmark: within $10 for the LLM proposer mean offer and within 2 percentage points for the LLM responder acceptance rate.

The rule was chosen for transparency and ease of comprehension. It gave participants a direct accuracy incentive: for any payment-relevant task and any realized benchmark, lowering the absolute prediction error weakly increases the chance that the report falls inside the rewarded tolerance interval. At the same time, the mechanism is not a proper scoring rule in the strict decision-theoretic sense. It should therefore be interpreted as incentivising accurate point predictions, rather than as eliciting full subjective probability distributions or exact subjective means. A formal discussion of the incentive properties and limitations of this rule is provided in Online Resource 1, Section S3.

4 Results

4.1 Aggregate belief miscalibration

Figure 1 presents a descriptive comparison between the large language model's behaviour in the Ultimatum Game and human participants' forecasts. The top panel shows the distribution of the 1,000 simulated proposer offers and, separately, the 108 participant point forecasts of the model's mean offer as a rug along the horizontal axis. The bottom-left panel compares the baseline observed responder acceptance frequencies with mean human forecasts across the six fixed offer amounts ($0, $10, $250, $500, $750, $1000). The bottom-right panel shows the cross-subject distribution of human responder forecasts at each offer level, with the corresponding baseline LLM acceptance frequency superimposed as a benchmark point. This presentation keeps the simulated behavioural outcomes and the elicited point forecasts conceptually distinct while showing their economically relevant level differences.

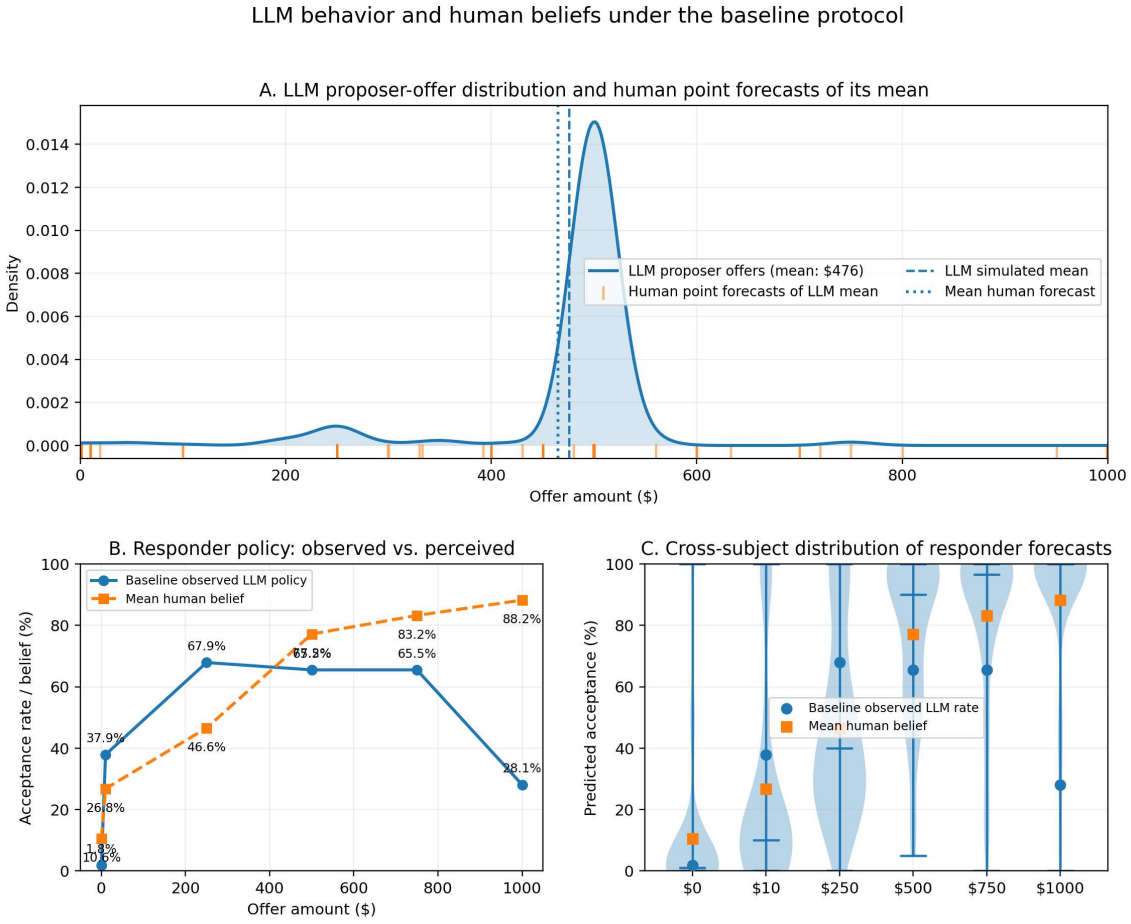


Figure 1: LLM behaviour and human beliefs under the baseline protocol. The proposer panel displays the simulated offer distribution and participant point forecasts of its mean as distinct objects; the responder panels compare baseline observed acceptance frequencies with the level and cross-subject distribution of human forecasts.

Alt text: Three-panel figure comparing the baseline LLM policy with human forecasts. The top panel shows a tightly concentrated LLM proposer-offer distribution near the equal split and participant point forecasts of the model mean. The lower panels compare observed LLM responder acceptance rates with mean and cross-subject human forecasts across six offers.

It is important to note that, for the responder task, the model is instructed to return only an integer response (1 = accept, 0 = reject). In practice, a subset of generations does not comply with this format. For all textual outputs from the LLM, we first manually reviewed and classified them into: (i) responses where the LLM provides the intended answer in textual form (e.g., "I accept" or "my answer is 1"), and (ii) entirely uninformative responses (e.g., refusals citing ethical concerns). These were subsequently categorized into three mutually exclusive classes: Accept (parseable as 1), Reject (parseable as 0), and Non-compliant (all remaining outputs). Throughout the paper, we report responder acceptance rates unconditionally: Accept is counted in the numerator and all 1,000 runs remain in the denominator, while Non-compliant remains a separate coded outcome. All baseline acceptance rates in the revised analysis are computed directly from the archived final post-classification dataset, so the reported benchmark is exactly reproducible from the replication file. Online Resource 1, Section S5 reports conditional acceptance rates

restricting the denominator to compliant 0/1 outputs across the temperature checks. The qualitative pattern is unchanged.

The top panel of Figure 1 juxtaposes two different statistical objects: the distribution of 1,000 simulated LLM offers and the cross-subject distribution of 108 point forecasts of the LLM's mean offer. The former is tightly concentrated near an even split, while the latter is heterogeneous across participants. Because participants were asked for a point forecast of the mean rather than for a distribution of possible LLM offers, we do not interpret differences in distributional shape as evidence about perceived model noisiness. The simulated mean offer is \$476 and the average participant forecast is \$465. A two-sided one-sample t-test of participant forecasts against the fixed simulated benchmark does not detect an aggregate mean bias ($t(107) = -0.43$, $p = 0.669$; 95% CI for the mean forecast: [\$413.9, \$516.1]).

Thus, the central proposer-side finding is not evidence of exact calibration or equivalence, but rather the absence of detectable aggregate mean bias together with substantial cross-subject heterogeneity in point forecasts. Under the baseline protocol, the model concentrates offers close to an equal split of the pie, in line with recent evidence that large language models can display human-like patterns of fairness and cooperation in classic games (Brookins and DeBacker, 2024; Lore and Heydari, 2024).

The responder-side results are different. Table 1 reports the LLM simulated acceptance rates and the average human beliefs for each offer amount. Two patterns emerge. First, participants systematically underpredict the LLM's willingness to accept low and intermediate offers. For offers of \$10 and \$250, human-predicted acceptance rates fall well below the model's simulated acceptance frequencies. This indicates that participants expect the LLM to behave more normatively—that is, to reject low or unfair offers more often—than it actually does under the baseline protocol. Second, participants substantially overestimate the LLM's acceptance of high offers, and this gap is especially large at the full-transfer offer of \$1000.

Read against the human-human benchmark discussed in the Introduction, these beliefs are broadly consistent with a human-like monotone acceptance heuristic: acceptance is expected to rise with the generosity of the offer, and very generous offers are expected to be accepted with high probability. This helps interpret why participants strongly overpredict acceptance at the full-transfer offer. Such a forecast is not surprising if subjects reason by analogy to human-human bargaining; it is surprising only relative to the observed LLM policy, which under the baseline protocol departs sharply from the monotone human-like pattern. The comparison should therefore be read as a belief-policy gap relative to the implemented artificial counterpart, rather than as evidence that participants would necessarily be well calibrated about human responders in the same task.

Table 1: LLM simulated acceptance rates and human beliefs across offer amounts

Offer amount	LLM (%)	Human (%)	Diff. (%)	95% CI Human	p-value	Sig.
\$0	1.8	10.6	+8.8	[6.4, 15.3]	$p < 0.001$	Yes
\$10	37.9	26.8	-11.1	[20.7, 33.3]	$p < 0.001$	Yes
\$250	67.9	46.6	-21.3	[41.2, 52.2]	$p < 0.001$	Yes
\$500	65.5	77.2	+11.7	[71.9, 82.2]	$p < 0.001$	Yes
\$750	65.5	83.2	+17.7	[78.1, 87.9]	$p < 0.001$	Yes
\$1000	28.1	88.2	+60.1	[82.9, 92.8]	$p < 0.001$	Yes

Notes: LLM acceptance rates are simulated frequencies based on 1,000 model runs per offer. Human acceptance rates are mean incentivised predictions across 108 participants. Confidence intervals refer to the cross-subject distribution of predictions. The p-values come from two-sided one-sample t-tests of participant forecasts against the fixed simulated LLM benchmark for each offer level.

The bottom-right panel of Figure 1 focuses on the cross-subject distribution of continuous human probability forecasts and superimposes the corresponding baseline acceptance frequency as a point benchmark. The economically relevant comparison is therefore between the observed policy level and the level and heterogeneity of beliefs about that policy, rather than between distributions of binary model realizations and continuous human forecasts.

From a game-theoretic perspective, the benchmark in the Ultimatum Game with monetary payoffs and fully rational selfish players is straightforward. The subgame perfect equilibrium has the proposer offer the smallest positive amount and the responder accept any positive transfer (Fudenberg and Tirole, 1991). Decades of laboratory evidence tell a very different story: human proposers commonly offer around forty to fifty percent of the stake, with equal splits as the modal choice and with low offers frequently rejected (Guth et al., 1982; Camerer, 2003; Oosterbeek et al., 2004). Against this backdrop, the behaviour of our large language model under the baseline proposer prompt sits close to the empirical equal-split regularity and far from the selfish equilibrium prediction. In our data, the LLM converges on offers just below one half of the pie with very limited dispersion. We interpret this as a prompt-conditioned policy that reproduces a familiar fairness convention, not as evidence that the model possesses a stable preference for fairness.

Participants, in turn, expect a different agent again. When asked to forecast the average transfer chosen by LLM proposers, the sample mean forecast is close to the simulated benchmark, but individual point predictions are heterogeneous. The theory sits at one extreme, while the model's baseline proposer behaviour and classic human data cluster near equal-split norms. The result should therefore be read as evidence of a belief-policy gap under the implemented protocol, not as evidence that the LLM has stable human-like preferences.

4.2 Individual-level miscalibration

Aggregate gaps do not identify who makes large mistakes, whether errors are mainly systematic or idiosyncratic, and whether miscalibration is concentrated at specific decision points. We therefore move to individual-level measures of miscalibration.

Proposer role. The LLM's simulated mean offer is $\mu_{AI} = 476$. Participant i predicts μ_i. We define the signed prediction error and the absolute prediction error as

$$e_i = \mu_i - \mu_{AI}, \tag{9}$$

$$|e_i| = |\mu_i - \mu_{AI}|. \tag{10}$$

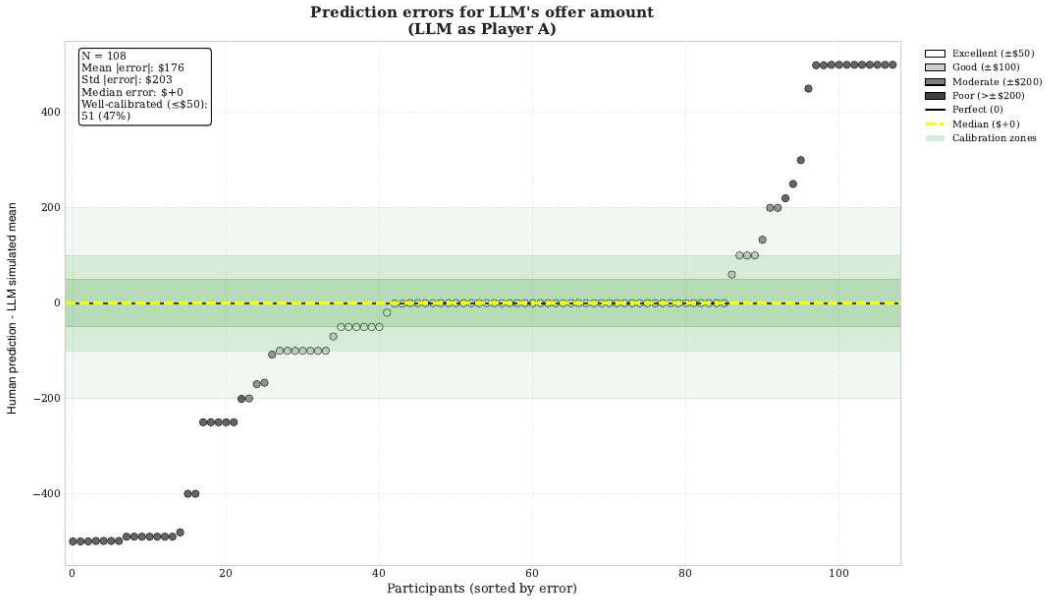


Figure 2: Human participants' miscalibration in predicting LLM offer.

Alt text: Ordered participant-level errors in predicting the LLM's mean proposer offer. Errors are centered near zero but spread widely across participants, showing substantial cross-subject heterogeneity in point forecasts.

Figure 2 reports individual-level prediction errors for the LLM offer when the LLM plays as proposer. Each point corresponds to one participant, ordered by signed error. The plot indicates a small aggregate bias but large heterogeneity in beliefs. The median error of zero confirms that the central belief is close to the simulated benchmark. At the same time, individual miscalibration is substantial: the mean absolute error is \$176 and the standard deviation of absolute errors is \$203. Only 47% of participants fall within the tight calibration band of \$50, and 58.3% fall within \$100. Thus, the absence of detectable aggregate mean bias coexists with substantial cross-subject heterogeneity in point forecasts.

Responder role. For each offer level $s \in \{1, \ldots, 6\}$, let q_s denote the LLM's simulated acceptance rate and let p_{is} be participant i's predicted acceptance rate. We compute prediction errors separately for each offer level:

$$e_{is} = p_{is} - q_s, \qquad |e_{is}| = |p_{is} - q_s|. \tag{11}$$

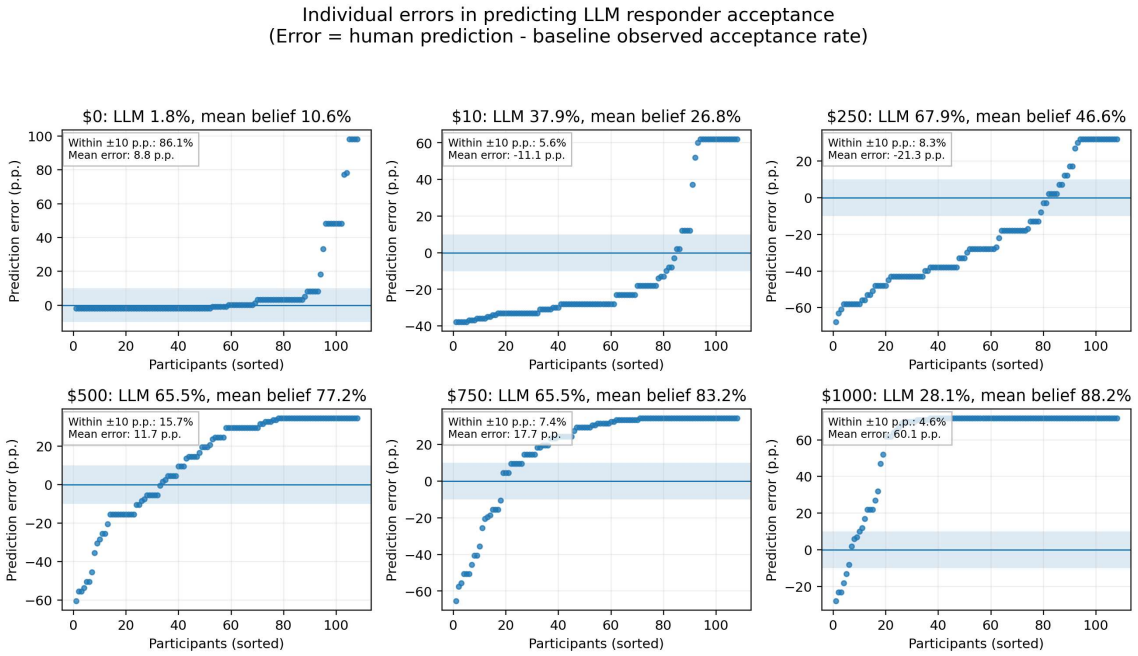


Figure 3: Human participant miscalibration in predicting LLM acceptance as responder.

Alt text: Six panels show ordered participant prediction errors for LLM responder acceptance at offers of 0, 10, 250, 500, 750, and 1000 dollars. Errors are mainly negative at low and intermediate offers and strongly positive at the full-transfer offer.

Figure 3 reports the distribution of participants' prediction errors in the responder role, separately for each offer level. Across panels, the figure reveals a pronounced and systematic pattern that is strongly offer-dependent. The zero-offer scenario is the best calibrated: 86.1% of participants predict the LLM's acceptance rate within 10 percentage points. At the low and intermediate offers ($10 and $250), participants exhibit substantial miscalibration: only 5.6% and 8.3%, respectively, fall within the well-calibrated band, and the distribution of signed errors is shifted toward negative values, indicating a tendency to underestimate the LLM's acceptance rates. At intermediate and high offers ($500 and $750), calibration remains limited, with only 15.7% and 7.4% of participants falling within the 10 percentage-point band. In sharp contrast, the $1000 panel shows a substantial upward shift: most participants overestimate acceptance at the highest offer, with only 4.6% of participants falling within the well-calibrated band.

Overall, the absence of detectable aggregate mean bias can mask substantial individual-level error. The sample mean proposer forecast is close to the simulated benchmark, but individual forecasts are highly dispersed across participants. In the responder role, errors are more systematic and strongly offer-dependent: participants tend to underestimate acceptance of low and intermediate offers and overestimate acceptance of high offers, especially the full-transfer offer.

4.3 Structure of individual beliefs

The six responder forecasts also allow us to examine whether participants' beliefs form coherent acceptance schedules rather than isolated point predictions. For participant i, we classify the elicited schedule as weakly monotone if $\hat{q}_i(x)$ is non-decreasing over the ordered offer grid $x \in \{0, 10, 250, 500, 750, 1000\}$. By this criterion, 92 of 108 participants

(85.2%) report a weakly non-decreasing acceptance schedule, and 43 (39.8%) report a strictly increasing schedule. Only 16 participants exhibit at least one downward step. This individual-level evidence supports the interpretation that most participants impose a familiar monotone acceptance logic on the artificial responder, even though the baseline observed LLM policy itself is non-monotone.

Beliefs across nearby offer levels are also strongly related. Pearson correlations between adjacent acceptance forecasts are $r = 0.508$ for \$0–\$10, $r = 0.865$ for \$10–\$250, $r = 0.552$ for \$250–\$500, $r = 0.737$ for \$500–\$750, and $r = 0.862$ for \$750–\$1000; all are statistically significant at $p < 0.001$. Thus, participants' responder forecasts display substantial internal structure rather than behaving like independent guesses at each offer level.

Beliefs also covary across the two LLM roles. Because we examine six proposer–responder correlations, we treat this analysis as exploratory and report Holm-adjusted p-values. Participants who predict a higher mean proposer offer tend to predict lower responder acceptance rates at \$10 ($r = -0.367$, $p_{\text{Holm}} < 0.001$) and \$250 ($r = -0.335$, $p_{\text{Holm}} = 0.002$). The associations at \$500 ($r = -0.231$, $p_{\text{Holm}} = 0.064$) and \$750 ($r = -0.197$, $p_{\text{Holm}} = 0.123$) do not remain significant at the 5% level after adjustment, and the corresponding correlations at \$0 ($r = -0.030$, $p_{\text{Holm}} = 0.758$) and \$1000 ($r = -0.093$, $p_{\text{Holm}} = 0.673$) are small. Descriptively, the two robust negative associations are consistent with some participants holding a more fairness-oriented mental model of the artificial counterpart—expecting it to be more generous as proposer and more selective as responder—although these exploratory correlations do not identify a specific belief-formation mechanism.

We next ask whether these structured belief-policy gaps are strategically consequential at the individual level.

4.4 Strategic consequences: aggregate and individual best responses

The belief gaps documented above are economically relevant because, in the ultimatum game, beliefs about the responder's acceptance policy directly determine the proposer's optimal offer. To quantify this implication, we use the framework developed in Section 2 and compute the expected payoff associated with each offer on the elicited grid. For each offer $x \in \{0, 10, 250, 500, 750, 1000\}$, expressed in dollars out of the \$1,000 stake, the expected payoff under the baseline observed LLM responder policy is

$$\pi(x) = (1000 - x)q(x),$$

where $q(x)$ is the simulated LLM acceptance frequency. The perceived expected payoff is instead

$$\hat{\pi}(x) = (1000 - x)\hat{q}(x),$$

where $\hat{q}(x)$ is the average human belief about the LLM's acceptance probability. This aggregate perceived policy is useful for visualization, but it is not equivalent to the distribution of subject-level best responses.

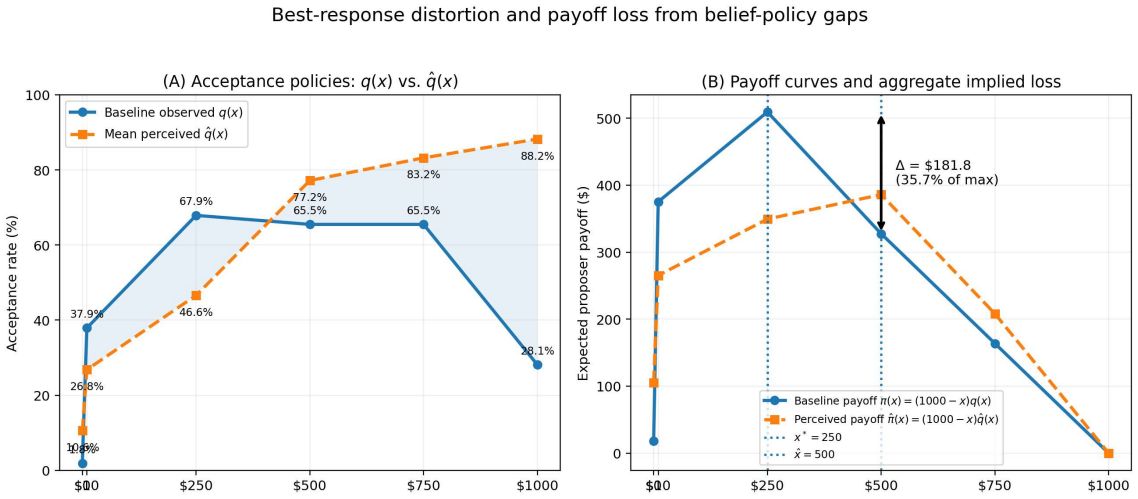


Figure 4: Best-response distortion and payoff loss from belief miscalibration. *Left panel*: baseline observed LLM acceptance policy $q(x)$ and average human-perceived policy $\hat{q}(x)$ across offer levels. *Right panel*: corresponding baseline-policy and perceived payoff curves $\pi(x)$ and $\hat{\pi}(x)$.

Alt text: Two-panel figure. The left panel compares the observed LLM acceptance policy with the average perceived policy across the offer grid. The right panel shows the corresponding proposer payoff curves: the observed-policy optimum is 250 dollars and the average-belief-implied optimum is 500 dollars.

This aggregate exercise shows that belief miscalibration induces a clear best-response distortion. Under the baseline observed LLM acceptance policy, the payoff-maximising offer on our grid is \$250. By contrast, under the average perceived acceptance policy, the payoff-maximising offer is \$500. In other words, a proposer who best-responds to the average human belief would offer an equal split, whereas the baseline observed LLM policy would make a substantially lower offer optimal.

Table 2: Best-response distortion and payoff loss from belief miscalibration

	Acceptance policy			*Expected payoff (\$ out of \$1,000 stake)*			
Offer x	$q(x)$ (%)	$\hat{q}(x)$ (%)	$\hat{q}(x) - q(x)$ (pp)	$\pi(x)$ (\$)	$\hat{\pi}(x)$ (\$)	$\hat{\pi}(x) - \pi(x)$ (\$)	Note
\$0	1.8	10.6	+8.8	18.0	105.6	+87.6	
\$10	37.9	26.8	−11.1	375.2	265.3	−109.9	
\$250	67.9	46.6	−21.3	509.3	349.4	−159.9	x^*
\$500	65.5	77.2	+11.7	327.5	385.9	+58.4	$\hat{x}$
\$750	65.5	83.2	+17.7	163.8	208.1	+44.3	
\$1000	28.1	88.2	+60.1	0.0	0.0	+0.0	

Notes to Table 2. $q(x)$ denotes the baseline observed LLM acceptance rate (unconditional; $N_{\text{sim}} = 1{,}000$) and $\hat{q}(x)$ denotes the average human-predicted acceptance rate ($N = 108$). $\pi(x) = (1000 - x)\, q(x)$ is the expected proposer payoff under the baseline observed policy in dollars out of the \$1,000 stake; $\hat{\pi}(x) = (1000 - x)\, \hat{q}(x)$ is the perceived expected payoff. The baseline-policy payoff-maximising offer is $x^* = \$250$, with $\pi(x^*) = \$509.25$; the belief-based best response is $\hat{x} = \$500$, with $\pi(\hat{x}) = \$327.5$. The payoff loss $\Delta = \pi(x^*) - \pi(\hat{x}) = \mathbf{\$181.75}$ ($\approx 35.7\%$ of the maximum achievable payoff) measures the

expected payoff foregone by a proposer who best-responds to the average perceived policy rather than the baseline observed policy. LLaMA-3.1-8B, $T = 0.8$.

The implied payoff consequences under counterfactual best responding are sizeable. Expressed in dollars out of the original \$1000 stake, the baseline-policy optimal offer of \$250 yields an expected proposer payoff of

$$(1000 - 250) \times 0.679 = 509.25.$$

By contrast, the belief-based best response of \$500 yields an expected payoff under the baseline observed policy of

$$(1000 - 500) \times 0.655 = 327.5.$$

Thus, a proposer who best-responds to the average perceived LLM acceptance policy would incur an expected payoff loss of

$$\Delta = 509.25 - 327.5 = 181.75.$$

This corresponds to \$181.75 out of the original \$1000 stake, or equivalently 18.18 cents per dollar of stake, about 35.7% of the maximum achievable payoff on the elicited grid.

Figure 4 illustrates this mechanism. The left panel compares the baseline observed LLM acceptance policy $q(x)$ with the average perceived policy $\hat{q}(x)$. The right panel plots the corresponding baseline-policy and perceived payoff curves. The vertical lines mark the baseline-policy payoff-maximising offer and the belief-based best response. The figure shows that participants' average beliefs imply systematic over-offering relative to the baseline observed LLM responder policy.

This result provides a direct empirical counterpart to Proposition 1. The estimated gap between $q(x)$ and $\hat{q}(x)$ is not a harmless forecasting error: it changes the best response implied by the perceived policy. In the present setting, a proposer who best-responded to the aggregate perceived policy would behave as if the LLM responder were more likely to reject unequal offers and more likely to accept equal or generous offers than it actually is. The resulting counterfactual best-response distortion takes the form of over-offering and would generate a substantial expected payoff loss under the baseline observed policy.

Individual belief-implied best responses. The aggregate calculation does not reveal whether the \$500 best response is an artefact of averaging heterogeneous beliefs. We therefore compute a best response separately for each participant. For participant i, let

$$\hat{\pi}_i(x) = (1000 - x)\hat{q}_i(x),$$

where $\hat{q}_i(x)$ is participant i's elicited acceptance forecast. Define the participant's maximizing set as

$$B_i = \arg\max_{x \in X} \hat{\pi}_i(x).$$

Of the 108 participants, 101 have a unique belief-implied best response and 7 have ties between two or more offers. Table 3 reports both the distribution of unique best responses and, for all participants, whether each offer belongs to the maximizing set.

Table 3: Distribution of individual belief-implied best responses

Offer	Unique best response, n (%)	Included in maximizing set, n (%)
\$0	6 (5.9)	7 (6.5)
\$10	21 (20.8)	21 (19.4)
\$250	13 (12.9)	16 (14.8)
\$500	60 (59.4)	67 (62.0)
\$750	1 (1.0)	5 (4.6)
\$1000	0 (0.0)	0 (0.0)

Notes: 101 of 108 participants have a unique belief-implied best response; 7 have ties. Percentages in the second column are calculated among the 101 participants with a unique best response. The third column uses all 108 participants and reports whether an offer belongs to the participant's maximizing set.

The individual-level results reinforce the aggregate finding. Among participants with a unique best response, \$500 is optimal for 60 of 101 (59.4%), while the baseline-policy optimum of \$250 is optimal for only 13 (12.9%). Looking at maximizing sets for all 108 participants, \$500 belongs to the maximizing set for 67 participants (62.0%), compared with 16 participants (14.8%) for \$250. Thus, the aggregate \$500 best response is not generated merely by averaging across otherwise unrelated individual belief profiles.

We next evaluate the payoff consequences of these individual best responses under the baseline observed LLM policy. For participants with a unique best response $\hat{x}_i$, define

$$\Delta_i = \pi(250) - \pi(\hat{x}_i),$$

where $\pi(250) = \$509.25$ is the maximum expected payoff on the elicited grid. Among the 101 participants with a unique best response, the mean payoff loss is \$168.4 and the median is \$181.8. For the seven participants with tied maximizing actions, we do not impose an arbitrary tie-breaking rule. Instead, evaluating all actions in each maximizing set yields an average-loss lower bound of \$164.3 and an upper bound of \$178.2 across the full sample. If the loss is averaged across tied maximizing actions for descriptive purposes, the full-sample mean is \$171.3 and the median remains \$181.8. Relative to the \$509.25 maximum attainable payoff, these losses amount to roughly one third of the attainable surplus.

The individual analysis therefore reaches the same substantive conclusion as the aggregate exercise while avoiding the aggregation concern: for a majority of participants, the elicited acceptance beliefs themselves imply an equal-split best response. If participants were to best-respond to those beliefs, the implied choices would generate economically meaningful payoff losses under the baseline observed policy. Belief-policy gaps about the observed artificial responder are therefore not only a matter of perception or trust, but a source of counterfactual payoff loss under belief-based optimization.

5 Protocol and Inference-Environment Sensitivity

The additional simulations are best viewed as an implementation-sensitivity exercise rather than as a conventional robustness check. The simulation archive now allows us to separate three distinct sources of variation. First, the Ollama baseline and the temperature-robustness runs all used exactly the user prompts reported in Online Resource 1, Section S1, with the same model parameters and infrastructure apart from the temperature setting.

These runs therefore identify sensitivity to decoding temperature within the documented Ollama environment. Second, an initial set of OpenRouter simulations used a shorter, single-paragraph prompt. Relative to the Online Resource 1, Section S1 baseline, this prompt was approximately one-third the length, omitted the social framing that the players would not knowingly meet, and explicitly stated Player A's retained payoff. We use this first-stage series only as evidence that the observed responder policy is sensitive to the implemented user prompt; because these three prompt features changed jointly, the exercise does not identify which individual wording change is responsible. Third, a second set of OpenRouter checkpoint simulations used the exact Online Resource 1, Section S1 baseline prompt, with only the stake and offer values substituted across conditions. This second-stage series provides the relevant comparison for inference-environment sensitivity because the user-level prompt is held fixed relative to the original Ollama baseline.

The OpenRouter checkpoint simulations were run in September 2026 using `meta-llama/llama-3.1-8b-instruct` through the OpenRouter API endpoint `https://openrouter.ai/api/v1`, with temperature $T = 0.8$ and 1,000 runs per cell at stakes of \$10, \$100, and \$1,000. Max tokens, top-p, and top-k were not explicitly specified and were therefore left at OpenRouter defaults. No seed was set. Each call used a newly constructed single-message input with no prior conversation history. No system prompt was explicitly supplied through the API; however, OpenRouter may apply provider-level default system instructions whose content is not exposed through the API. This qualification is part of the inference environment itself and is one reason we interpret the comparison at the implementation-environment level rather than as a clean estimate of a single provider-specific mechanism.

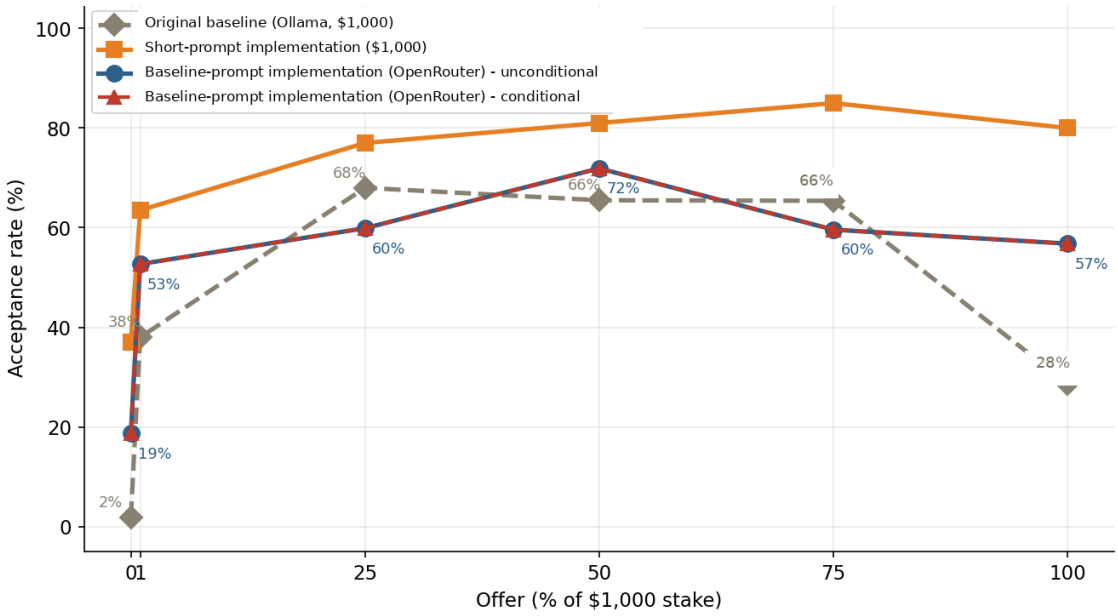


Figure 5: Protocol and inference-environment sensitivity at a fixed $1,000 stake. The figure compares the original Ollama baseline, the first-stage OpenRouter short-prompt implementation, and the second-stage OpenRouter checkpoint replication using the exact Online Resource 1, Section S1 baseline user prompt. The short-prompt implementation differed from the baseline in three documented respects—length, social framing, and explicit statement of Player A's retained payoff—and is used descriptively to assess prompt-implementation sensitivity. The exact-prompt OpenRouter checkpoint series is the relevant inference-environment comparison. Unconditional acceptance uses all 1,000 runs in the denominator; conditional acceptance restricts the denominator to parseable 0/1 outputs.

Alt text: Acceptance-rate curves at a 1,000-dollar stake compare the original Ollama baseline, a first-stage OpenRouter short-prompt implementation, and second-stage OpenRouter simulations using the exact baseline prompt. The exact-prompt OpenRouter curve differs from the Ollama baseline despite matching the named model, prompt, and temperature.

At a fixed $1,000 stake, the first-stage short-prompt series differs substantially from the simulations using the full Online Resource 1, Section S1 prompt over much of the offer grid. This confirms that the observable policy can change when the user-level description of the same bargaining problem changes. We interpret this comparison at the level of prompt implementation rather than as an estimate of any single wording effect, because the shorter prompt simultaneously changes length, social framing, and the explicit description of Player A's payoff.

The second-stage OpenRouter checkpoint simulations allow a cleaner comparison of inference environments. They use the same Online Resource 1, Section S1 user prompt, the same temperature ($T = 0.8$), the same $1,000 stake, and the same number of runs per cell as the Ollama baseline, but are served through OpenRouter rather than the local Ollama environment. The resulting policy does not reproduce the original Ollama baseline. The clearest discrepancy occurs at the full-transfer offer: the original baseline acceptance rate is 28.1%, whereas the OpenRouter unconditional acceptance rate is 56.8%. Differences are also visible at other offer levels. Because provider-level default system instructions, serving

configurations, model builds, routing choices, and other backend details are not fully observable, we do not attribute this divergence to a single hidden mechanism. Instead, it provides evidence that the observed policy is sensitive to the broader inference environment even when the user-level prompt and the main user-controlled decoding parameter are held fixed.

For readability, the full cross-stake diagnostic is reported in Online Resource 1, Section S4 (Figure S1).

The cross-stake results in Online Resource 1, Figure S1 show why a simple stake-size invariance interpretation is not warranted. In the \$10 condition, the baseline prompt produces non-compliant outputs in nearly all runs for positive offers, so unconditional acceptance mechanically collapses toward zero and conditional rates are generally not defined because there are too few parseable 0/1 responses. A similar problem occurs in the \$100 condition at the zero offer. These cells therefore do not provide clean evidence about how acceptance changes with stake size; instead, they reveal an interaction between the numerical scale of the task and the model's compliance with the requested response format.

Where compliance is sufficiently high, conditional and unconditional acceptance rates are generally close, indicating that the main differences across interpretable cells are not driven solely by the coding of non-compliant outputs. But the small-stake failures are themselves substantively relevant: the behavioral object implemented by an LLM includes not only accept/reject frequencies conditional on compliance, but also whether the system follows the requested action format at all.

Taken together, the additional simulations sharpen the protocol-contingency interpretation developed in Section 2. The documented Ollama runs show sensitivity to decoding temperature; the first-stage OpenRouter exercise shows sensitivity to prompt implementation; the cross-stake checkpoint simulations reveal sensitivity to numerical scale and response-format compliance; and the exact-prompt Ollama–OpenRouter comparison shows sensitivity to the broader inference environment. These exercises are complementary rather than interchangeable: only the checkpoint comparison holds the user-level prompt fixed across serving environments, while the short-prompt exercise deliberately changes the prompt. Accordingly, the original Ollama $T = 0.8$ simulations remain the relevant benchmark for the human-belief analysis because they are the simulations participants were asked to predict. The sensitivity results instead delimit the external validity of that benchmark: the estimated belief-policy gap is a property of a specified artificial-agent implementation and should not be interpreted as a stable feature of LLaMA-3.1-8B across prompts, temperatures, numerical representations, or serving environments.

6 Discussion

Building on these findings, belief-policy gaps in AI-mediated bargaining may have tangible consequences in several applied domains, echoing broader evidence that people systematically misjudge and miscalibrate their trust in algorithmic systems (Dietvorst et al., 2015, 2018). Importantly, participants' forecasts diverge substantially from the observed LLM acceptance policy, with no corresponding aggregate gap in their expectations of proposer mean offers. The individual-level evidence clarifies the structure of this gap: 85.2% of participants report a weakly monotone perceived acceptance schedule, and beliefs at adjacent offer levels are strongly correlated. This is consistent with most participants

relying on a coherent human-like monotonic heuristic—higher offers should be accepted at least as often as lower ones—whereas the observed LLM acceptance policy under our baseline protocol is non-monotone, most visibly at the full-transfer offer. In the Ultimatum Game, such differences matter because beliefs about acceptance probabilities directly discipline the proposer's optimal offer.

An advantage of the beliefs-as-policy approach is precisely that it does not require attributing a latent utility function to the artificial counterpart. The LLM was given the rules and consequences of the game but no experimenter-imposed monetary objective. Our object of interest is therefore the mapping from the specified protocol to observed behaviour, and the extent to which human subjects correctly perceive that mapping. The payoff-maximisation problem enters only when evaluating a human proposer's best response to the observed or perceived acceptance policy. This separation between artificial policy and human optimisation makes the strategic cost of misperception identifiable without assuming that the LLM itself behaves as a money-maximising economic agent.

Because we do not elicit beliefs about a human responder in the same experimental design, this interpretation should be read as contextual rather than causal. The extensive human-human Ultimatum Game literature motivates the human-like monotone acceptance benchmark, but our data do not identify AI-specific miscalibration in the strongest sense. Put differently, the paper shows that participants' beliefs are poorly calibrated to the LLM policy observed under our protocol; it does not show that the same participants would have been perfectly calibrated to a human responder. A matched human-responder belief treatment would be needed to make that stronger comparison.

The strategic consequences are also visible at the subject level. The equal-split offer of \$500 is not merely the best response to the average perceived policy: it is the unique belief-implied optimum for 60 of the 101 participants with a unique optimum and belongs to the maximizing set for 62.0% of the full sample. The individual belief-implied best responses therefore reproduce the aggregate distortion rather than averaging it into existence. Under counterfactual best responding, the associated payoff losses would be economically substantial: among unique best responses, the mean loss is \$168.4 and the median is \$181.8 relative to the \$509.25 maximum attainable payoff on the elicited grid. Projecting a human-like acceptance pattern onto the model—or otherwise misperceiving its behavioral policy—can therefore imply a quantifiable expected payoff loss.

The sensitivity analysis adds an important qualification to this interpretation while also separating distinct implementation channels. The Ollama runs document temperature sensitivity under a fixed prompt and serving environment. The first-stage OpenRouter exercise shows that changing the user prompt can materially alter the responder policy, although the simultaneous changes in prompt length, social framing, and payoff description prevent attribution to any single wording feature. Most importantly for portability, the second-stage OpenRouter checkpoint simulations use the exact Online Resource 1, Section S1 user prompt and $T = 0.8$ yet still differ from the original Ollama baseline. Together with the cross-stake non-compliance patterns, these results imply that measured LLM behavior in strategic tasks should be treated as conditional on the full implementation protocol—including prompt, decoding parameters, numerical representation, model serving, and inference environment—rather than as a stable behavioral trait of the named model.

In many applied bargaining environments, the relevant problem is not only what an artificial agent does, but whether humans correctly understand the acceptance policy they face. Our results speak directly to this issue. If users hold inaccurate beliefs about the probability that an artificial counterpart will accept different offers, they may choose

offers that are too generous, too aggressive, or otherwise poorly matched to the observed policy of the system.

A first example is consumer dispute resolution. AI-supported mediation and online dispute resolution systems are increasingly discussed for small-value claims and mass consumer disputes (Alessa, 2022; Paterson et al., 2025; Sharma et al., 2025). In such settings, a consumer who believes that an artificial agent will behave like a conciliatory human mediator may make a low initial demand or accept a modest settlement too quickly. Conversely, a consumer who overestimates the agent's willingness to accept extreme claims may make offers that are repeatedly rejected, increasing delay, frustration, and transaction costs. The central mechanism is the same as in our experiment: mistaken beliefs about the responder's acceptance policy distort bargaining behaviour.

A second domain is wage bargaining in digital labour markets. Firms increasingly rely on algorithmic management and AI-assisted tools to screen applicants, rank workers, filter counteroffers, or generate salary recommendations (Milanez et al., 2025; Faioli, 2025). When workers face an opaque algorithmic or AI-mediated bargaining process, they must form beliefs about what kinds of wage requests or counteroffers the system is likely to accept. If they underestimate acceptance probabilities, they may concede too early and leave surplus on the table. If they overestimate acceptance probabilities, they may make demands that are unlikely to be accepted, potentially reducing the probability of agreement. In both cases, belief miscalibration can shape bargaining outcomes even when the underlying allocation rule is fixed.

Related mechanisms arise in debt restructuring, insurance claims, and tax settlements. These are environments in which individuals or firms often make proposals under uncertainty about the acceptance rule used by an institution or platform. If an automated or AI-assisted system evaluates repayment plans, compensation claims, or settlement offers, users must infer the system's acceptance policy from limited information. Holding inaccurate beliefs about that policy may lead to inefficient proposals: offers that are unnecessarily generous, offers that are too aggressive and likely to be rejected, or repeated bargaining delays. Thus, belief-policy gaps in artificial acceptance policies can generate welfare losses not because the artificial system is irrational, but because humans best-respond to the wrong perceived policy.

These examples suggest that belief calibration should be treated as a design and governance issue. Transparency should not only disclose that an artificial agent is being used, but also help users form accurate expectations about its relevant behavioural policy. In bargaining environments, this may require simple summaries of historical acceptance rates, explanations of the main constraints guiding the system, or interfaces that allow users to explore likely outcomes before committing to an offer. Such tools could reduce payoff losses by aligning perceived acceptance policies more closely with actual ones.

These examples indicate that regulators and designers should concern themselves not only with what AI agents actually do but also with what people believe they do. Work on algorithmic transparency and accountability argues that clear disclosure of system purpose, constraints and governance is central to trustworthy deployment (Kossow et al., 2021; Cheong, 2024; Park and Yoon, 2025), while the literature on explainable and transparent AI highlights that simple summaries of past decisions and outcomes can help users form more accurate mental models (Guidotti et al., 2019; Valderrama et al., 2023; Ning et al., 2024). Experimental evidence on algorithm aversion further suggests that allowing users limited control or the ability to adjust algorithmic outputs can mitigate distrust and improve uptake (Dietvorst et al., 2018). In our context, improving transparency about

both proposer offer patterns and responder acceptance rates—e.g., interfaces that display historical acceptance rates for different offers and interactive tools that let users experiment before committing—are natural candidates to align expectations with reality and reduce the costs associated with miscalibrated beliefs about artificial policies.

7 Conclusion

This study examines belief formation about an artificial bargaining counterpart under a fixed interaction protocol in a canonical bargaining environment. In our Ultimatum Game, the large language model's proposer policy resembles a familiar empirical regularity from human laboratory play: offers are concentrated close to an equal split of the pie (Guth et al., 1982; Camerer, 2003; Oosterbeek et al., 2004). However, this similarity is incomplete. The model's proposer behaviour is much less dispersed than the cross-subject distribution of participants' point forecasts, and its responder policy is not well described by a simple human-like monotonic acceptance rule. Acceptance varies substantially across offer levels and departs sharply from participants' expectations, especially at the extremes.

The main finding is not that participants are generally poor at predicting LLMs. Rather, under our protocol, their forecasts differ systematically from the observed policy of the artificial counterpart. For the proposer role, we detect no aggregate mean bias relative to the simulated benchmark, but point forecasts display substantial cross-subject heterogeneity. For the responder role, belief-policy gaps are systematic and directionally biased: participants underestimate acceptance of some low and intermediate transfers and overestimate acceptance of high transfers, especially the full-transfer offer. At the same time, these beliefs are internally structured: 92 of 108 participants (85.2%) report a weakly non-decreasing perceived acceptance schedule, consistent with a human-like monotone acceptance heuristic that does not describe the baseline observed LLM policy.

Crucially, these belief-policy gaps are payoff-relevant at the individual level. The baseline observed LLM responder policy makes \$250 optimal on the elicited grid, whereas \$500 is the unique belief-implied best response for 60 of the 101 participants with a unique optimum and belongs to the maximizing set for 67 of all 108 participants. The aggregate equal-split best response is therefore not an artefact of averaging heterogeneous beliefs. Among participants with a unique best response, the mean payoff loss is \$168.4 and the median is \$181.8 relative to the \$509.25 maximum attainable payoff. Misperceiving the model's behavioral policy is therefore not merely a forecasting error: when translated into a best response, it can generate a quantifiable expected payoff loss.

The comparison between theoretical equilibrium (Fudenberg and Tirole, 1991), human laboratory regularities (Guth et al., 1982; Camerer, 2003), observed large language model play (Brookins and DeBacker, 2024; Lore and Heydari, 2024), and human expectations reveals a behavioural wedge. The LLM produces superficially equitable proposer offers, yet its responder policy departs from the monotonic human-like acceptance pattern participants appear to expect. Because we do not include a contemporaneous human-responder belief treatment, this comparison should be interpreted as contextual rather than causal: the human-human literature motivates the hypothesis that participants may rely on a monotone acceptance heuristic, while our data show that this heuristic is poorly calibrated to the LLM policy observed under our protocol. The relevant contribution is therefore not a definitive test of AI-specific miscalibration, but the documentation of a payoff-relevant belief-policy gap for a specific artificial bargaining counterpart. The

sensitivity analysis in Section 5 further shows that the observed responder policy changes with prompt formulation, numerical scale, and inference environment. The baseline should therefore be interpreted as a protocol-specific policy rather than as evidence of a stable underlying preference pattern of the named model.

These findings have implications for the emerging use of artificial agents in negotiation, dispute resolution, platform-mediated bargaining, and other strategic settings. When humans bargain with or through artificial agents, they must form beliefs about the policy such agents implement. If those beliefs are miscalibrated, agents may make offers that are too generous, too aggressive, or otherwise poorly matched to the observed acceptance policy they face. Misalignment in expectations about acceptance probabilities is especially relevant in contexts where offers are made infrequently, under time pressure, or with limited opportunities for learning (Alessa, 2022; Paterson et al., 2025; Sharma et al., 2025). In this sense, belief-policy gaps about artificial acceptance policies create an additional layer of friction on top of standard concerns about transparency and accountability (Kossow et al., 2021; Cheong, 2024; Park and Yoon, 2025).

At the same time, our design has clear limitations. We study one canonical game, one stake structure, one large language model, and a relatively homogeneous student sample. The results should therefore be read as a controlled measurement of beliefs about the observed policy of a specific artificial counterpart under a fixed protocol, not as a general characterization of all LLM behaviour across strategic environments. A further limitation is that we do not elicit, within the same experimental design, participants' beliefs about a human responder. For this reason, our data do not identify AI-specific miscalibration in the strongest sense. Instead, the human-human Ultimatum Game literature provides a contextual benchmark against which to interpret the belief patterns we observe. Future work should extend the analysis to other games, richer dynamic settings, alternative models and decoding protocols (Fan et al., 2024; Lore and Heydari, 2024; Huang et al., 2025), and should compare beliefs about artificial counterparts with beliefs about human counterparts in matched designs.

A natural next step is to link beliefs about artificial play to actual incentivised choices when humans bargain with or through these systems. Another important direction is to test whether simple information tools—such as disclosure about the agent's objectives, summaries of past behaviour, or interfaces displaying historical acceptance rates—can improve belief calibration and reduce payoff losses (Guidotti et al., 2019; Ning et al., 2024; Dietvorst et al., 2018). More broadly, the results suggest that the strategic role of artificial agents depends not only on how such agents behave, but also on whether humans understand the behavioural policies they are facing.

Statements and Declarations

Competing interests. The authors declare no competing interests.
Data availability. An anonymized replication package containing the analysis data, simulation metadata, prompts, and reproducibility code is supplied as Online Resource 2. The public archival repository will be provided in the final published version.
AI-assisted tools. During manuscript revision, an AI language model was used to assist with language editing, restructuring, drafting of revisions, and verification of selected calculations from author-provided data. All AI-assisted outputs were critically reviewed and edited by the authors, who take full responsibility for the final content and reported

results.

References

Aher, G. V., Arriaga, R. I., and Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202, 337–371.

Alessa, A. (2022). The role of artificial intelligence in online dispute resolution: A brief and critical overview. *Information & Communications Technology Law*, 31(3):319–342.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Azar, O. H., Lahav, Y., and Voslinsky, A. (2015). Beliefs and social behavior in a multi-period ultimatum game. *Frontiers in Behavioral Neuroscience*, 9:29.

Barak, A. and Costa-Gomes, M. (2025). Humans expect rationality and cooperation from LLM opponents in strategic games. arXiv preprint arXiv:2505.11011.

Brookins, P. and DeBacker, J. M. (2024). Playing games with GPT: what can we learn about a large language model from canonical strategic games? *Economics Bulletin*, 44(1):25–37.

Buchanan, J. and Hickman, W. (2024). Do people trust humans more than ChatGPT? *Journal of Behavioral and Experimental Economics*, 112:102239.

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

Cheong, B. C. (2024). Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics*, 6:1421273.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170.

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, 14:583–610.

Faioli, M. (2025). Digitalisation, artificial intelligence and algorithmic management in the workplace: Shaping the future of work. Available at SSRN 5665734.

Fan, C., Chen, J., Jin, Y., and He, H. (2024). Can large language models serve as rational players in game theory? A systematic analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:17960–17967.

Forsythe, R., Horowitz, J. L., Savin, N. E., and Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6(3):347–369.

Fudenberg, D. and Tirole, J. (1991). *Game Theory*. MIT Press.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):93.

Guth, W., Schmittberger, R., and Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4):367–388.

Huang, J.-t., Li, E. J., Lam, M. H., Liang, T., Wang, W., Yuan, Y., Jiao, W., Wang, X., Tu, Z., and Lyu, M. (2025). Competing large language models in multi-agent gaming environments. In *The Thirteenth International Conference on Learning Representations*.

Kossow, N., Windwehr, S., and Jenkins, M. (2021). Algorithmic transparency and accountability. Anti-corruption helpdesk report, Transparency International.

Lore, N. and Heydari, B. (2024). Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14:18490.

Milanez, A., Lemmens, A., and Ruggiu, C. (2025). Algorithmic management in the workplace: New evidence from an OECD employer survey. OECD Artificial Intelligence Papers.

Ning, X., Lu, Y., Li, W., and Gupta, S. (2024). How transparency affects algorithmic advice utilization: The mediating roles of trusting beliefs. *Decision Support Systems*, 183:114273.

Oosterbeek, H., Sloof, R., and van de Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, 7(2):171–188.

Valderrama, M., Hermosilla, M. P., and Garrido, R. (2023). State of the evidence: Algorithmic transparency. Open Government Partnership, 24.

Ozkes, A. I., Hanaki, N., Vanderelst, D., and Willems, J. (2024). Ultimatum bargaining: Algorithms vs. humans. *Economics Letters*, 244:111979.

Park, K. and Yoon, H. Y. (2025). AI algorithm transparency, pipelines for trust not prisms: Mitigating general negative attitudes and enhancing trust toward AI. *Humanities and Social Sciences Communications*, 12:1160.

Park, J. S., Popowski, L., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2022). Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*.

Paterson, J. M., Tan, V., and Webb, J. (2025). Generative AI in small value consumer disputes: Reviving not resolving challenges of design and governance in online dispute resolution. Available at SSRN 5313046.

Sharma, S., Mishra, Y. R., Sharma, R., Kumar, R., Ranjan, R. C., and Kumari, S. (2025). Artificial intelligence and online dispute resolution in consumer tribunals. *Advances in Consumer Research*, 2:4561–4567.

Trautmann, S. T. and van de Kuilen, G. (2015). Belief elicitation: A horse race among truth serums. *The Economic Journal*, 125(589):2116–2135.

Vavra, P., Chang, L. J., and Sanfey, A. G. (2018). Expectations in the Ultimatum Game: Distinct effects of mean and variance of expected offers. *Frontiers in Psychology*, 9:992.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MindtheGapJEICOnlineResource1SupplementaryInformation.pdf