Large Language Models can Predict Human Strategic Decisions*

Pedro Gonzalez-Fernandez[1], Siting Estee Lu[2], and Helena Normann[3]

[1]Heidelberg Universisty
[2]Edinburgh University
[3]Paris School of Economics

January 17, 2026

Abstract

We study whether large language models (LLMs) can predict human strategic behavior from pre-play communication. Using three canonical laboratory games that vary in incentive alignment and communication structure, we provide LLMs and incentivized human forecasters with identical transcripts and ask them to predict players' subsequent actions. GPT-5 consistently outperforms humans and achieves accuracy well above chance, especially when incentives are aligned and communication is bilateral. The performance gap arises almost entirely from correctly forecasting cooperative actions, while both humans and GPT–5 struggle to anticipate defection. These results suggest that strategic communication contains systematic information about future behavior that humans underutilize and that LLMs are able to exploit more effectively.

Keywords Large Language Models · Strategic Communication · Behavior Prediction

JEL Classification D83 · C92 · C88

*We are very thankful to Tim Cason for sharing his experimental data. For helpful comments and discussions, we thank Peter Duersch, Julian Matthes, Joerg Oechssler, Thomas Meissner, Christoph Vanberg, Roy Zultan, participants at the BEELab seminar at Maastricht University, and participants at the 21st HeiKaMax workshop. For a useful prompting workshop we thank Can Celebi and Stefan Penczynski. Gonzalez-Fernandez acknowledges funding from Maastricht Graduate School of Business and Economics. All data can be found here. The LLM implementation described in the paper is openly accessible on GitHub. Corresponding author: Pedro Gonzalez-Fernandez (pedro.gonzalez-fernandez@awi.uni-heidelberg.de)

1 Introduction

Communication plays a central role in many economic interactions, from coordination and bargaining to trust and collective decision-making. A large theoretical literature studies how costless messages can shape beliefs and behavior, beginning with the seminal cheap talk model of Crawford and Sobel (1982) and extending to richer analyses of communication equilibria and credibility (Farrell and Rabin, 1996). A core insight of this literature is that communication may be informative even when it is not directly payoff-relevant, but that its informativeness depends critically on incentive alignment and strategic structure (e.g., Agranov, 2025).

Laboratory experiments provide mixed evidence on the extent to which communication predicts subsequent actions. In some environments, pre-play messages facilitate coordination and cooperation (Cooper et al., 1989; Charness and Dufwenberg, 2006; Fonseca and Normann, 2012). In others, communication is noisy, strategically misleading, or weakly correlated with behavior, reflecting incentives to deceive or to engage in non-binding reassurance (Gneezy, 2005; Vanberg, 2008). As a result, it remains unclear how much predictive content is contained in actual messages once strategic incentives are taken into account.

Despite extensive theoretical and experimental work on communication, a basic empirical question remains largely unexplored: given observed communication, to what extent can one forecast the actions that players will ultimately take? This question is nontrivial. A large body of evidence suggests that individuals have difficulty inferring others' intentions from communication in strategic settings, even when payoff structures are transparent. Humans perform only slightly above chance in detecting deception from verbal cues (Bond and DePaulo, 2006), and stated beliefs and subsequent play are often inconsistent even with fully transparent payoffs (Costa-Gomes and Weizsäcker, 2008). At the same time, recent advances in natural language processing have demonstrated that modern algorithms can extract subtle and high-dimensional regularities from unstructured text at scale (Brown et al., 2020). This raises the possibility that human-to-human communication may contain predictive signals about strategic behavior that are difficult for us humans to identify consistently, but could nonetheless be detected by algorithms.

In this paper, we evaluate whether large language models (LLMs) can forecast human strategic behavior using pre-play communication. We focus on settings in which players exchange natural-language messages before making payoff-relevant decisions, and ask whether the content of these messages can be used to anticipate subsequent actions. Our analysis centers on GPT–5, alongside earlier-generation models. Their predictive performance is benchmarked against incentivized human forecasters who observe the same communication transcripts.

We study three canonical experimental environments drawn from the literature on pre-play communication: Minimum Effort Game and Prisoner's Dilemma (experimental data from Cason et al. (2019)) and Trust (Entry) Game (experimental data from Charness and Dufwenberg (2006)). These environments differ systematically along two dimensions: (i) whether incentives are aligned or conflicting, and (ii) whether communication is bilateral or unilateral. Together, these features generate clear differences in how informative language is expected to be about subsequent actions. When incentives are aligned and communication is two-way, messages should more reliably reflect intentions; when incentives conflict and communication is one-way, messages may be strategically misleading and less informative.

This structure allows us to formulate testable hypotheses consistent with the literature on communication structure in games.[1] If pre-play messages contain predictive information about behavior, forecasting accuracy should vary systematically across environments in line with their expected informativeness. Conversely, if language is largely uninformative or strategically noisy, predictive performance should be limited and invariant across settings. By comparing model and human forecasts across these environments, we provide a principled test of whether (and when) communication reveals actionable information about strategic behavior.

Our empirical strategy is straightforward. We use LLMs to generate out-of-sample forecasts of players' actions based on the content of pre-play communication. To benchmark these predictions, we run a new incentivized forecasting experiment in which human subjects observe the same transcripts and face the same binary prediction tasks. To ensure comparability, we maintain informational symmetry between LLMs and humans. We then assess models' predictive performance relative to chance, against human forecasters, across strategic environments, as well as across models. This design allows us to assess whether language contains systematic predictive signals and whether such signals are exploited differently by humans and by LLMs.

Our main findings can be summarized as follows. First, GPT–5 predicts strategic behavior well above chance in all three environments. Second, GPT–5 significantly outperforms incentivized human forecasters in each setting. Third, predictive accuracy varies systematically across environments exactly in line with expected informativeness: performance is highest in settings with two-way communication and aligned incentives, and lowest in settings with one-way communication and conflicting incentives. Fourth, GPT–5's forecasts cannot be explained by simple base-rate imitation; prediction accuracy reflects systematic sensitivity to conversational content. Finally, the ability to forecast strategic behavior is not unique to GPT–5: even earlier-generation and lightweight language models perform at least as well as human forecasters across all environments.

A closer look at conditional performance also reveals an important asymmetry in how predictive accuracy is achieved. GPT–5's high accuracy is driven primarily by its ability to correctly identify cooperative outcomes, where it substantially outperforms human forecasters in all three environments. By contrast, accuracy conditional on defection is markedly lower and not significantly different from that of humans. Thus, while LLMs reliably detect signals associated with cooperative behavior, predicting defection remains difficult for both humans and machines.

Lastly, we present some descriptive statistics for the "superforecasters", who outperform GPT–5. There is a small portion of individuals who exceeds GPT–5's mean accuracy in Prisoner's Dilemma and Trust (Entry) Game, suggesting individuals may do better in environments with conflicting incentives.

Literature. A large experimental literature studies pre-play communication across a variety of strategic environments, documenting that the relationship between talk and

[1] Agranov (2025) synthesized diverse experimental evidence, emphasizing that communication effects are highly sensitive to strategic structure, incentive alignment, and design features such as message richness and group size.

subsequent behavior is highly context dependent. In coordination games and related settings with Pareto-ranked equilibria, communication is often associated with improved efficiency and equilibrium selection (Moreno and Wooders, 1998; Blume and Ortmann, 2007; Ellingsen and Östling, 2010). In social dilemmas, repeated interaction, and team or group environments, communication has been shown to affect cooperation, collusion, and joint outcomes, though with substantial heterogeneity across designs (Sutter and Strassmair, 2009; Palfrey et al., 2017; Proto et al., 2017; Sibly and Tisdell, 2018; Cooper and Kagel, 2023). Related evidence from bargaining, trust, and contract-like settings similarly finds that messages can influence behavior, but through channels that mix strategic incentives with social and psychological considerations (Charness and Dufwenberg, 2006; Ben-Ner and Putterman, 2009; Ben-Ner et al., 2011; Agranov and Tergiman, 2019; Zultan, 2012; Greiner et al., 2014). Taken together, this work establishes that communication can matter, but also that its behavioral implications vary widely across environments and cannot be summarized by a single effect.

A complementary strand of work emphasizes the limits of communication as a predictor of behavior when incentives conflict or strategic uncertainty is salient. Experimental studies contrast cheap talk with alternative information channels and show that words alone may be weakly informative when misrepresentation is profitable (Duffy and Feltovich, 2002). Others analyze how communication interacts with actions and reputational concerns when multiple signals are available (Duffy and Feltovich, 2006; Meloso et al., 2023), or how strategic uncertainty constrains the effectiveness of talk in bargaining and coordination contexts (Burton and Sefton, 2004; Feltovich and Swierzbinski, 2011). Additional evidence highlights that deceptive or uninformative communication need not take the form of explicit lies, but may involve silence, selective disclosure, or evasive language (Sánchez-Pagés and Vorsatz, 2009; Alempaki et al., 2024), and that observed behavior may be better explained by bounded strategic reasoning than by equilibrium refinements alone (Kawagoe and Takizawa, 2009). While this literature clarifies when and why communication may fail, it typically evaluates outcomes within individual games and does not directly assess whether realized messages contain exploitable signals for forecasting subsequent actions across environments (Backus et al., 2019; Li et al., 2022; Charness et al., 2023).

Finally, we relate to a growing literature that uses large language models (LLMs) in economics and computational social science. Existing work commonly studies LLMs as simulated decision-makers in economic environments (Horton, 2023), examines their behavior in repeated strategic interactions (Akata et al., 2025), evaluates their rationality and strategic reasoning ability within game-theoretic frameworks (Guo et al., 2024), and benchmarks their performance on formal reasoning tasks (Duan et al., 2024). Another strand applies LLMs to classification and auditing problems, including persuasion and deception-related labeling (Apel et al., 2022; Rogiers et al., 2024; Loconte et al., 2023; Hazra and Majumder, 2024; Kretschmar et al., 2025; Shahriar, 2025) and applied detection tasks such as financial misrepresentation (Erva Ergun and Sefer, 2025; Kirkos et al., 2024). These approaches typically rely on curated labels, synthetic prompts, or LLMs acting as agents, and therefore abstract from the strategic uncertainty inherent in pre-play communication, where the mapping from language to actions is neither deterministic nor reducible to simple lie detection (Alempaki et al., 2024; Meloso et al., 2023). We instead use LLMs as out-of-sample forecasting devices for real human strategic decisions based on naturally occurring communication, and benchmark their predictive performance directly against incentivized human forecasters, providing a measure of when and to what extent language predicts action across strategic environments (Luo et al., 2025).

The rest of the paper is structured as follows: section 2 introduces our taxonomy to classify the strategic environments and introduces the games we analyze, section 3 explains the forecasting procedure for LLMs and our experimental design for human forecasters. Section 4 presents our results, and section 5 discusses future research avenues and concludes.

2 Strategic Settings

2.1 Conceptual Overview

We study three strategic environments that differ in how communication relates to economic incentives, and in turn, in how informative language is expected to be. To achieve this, we selected three different games with pre-play communication from previously run experiments. Our goal is not to re-analyze these three isolated cases, but to use them as illustrations of communication environments that differ along two key dimensions: (i) whether incentives make messages more likely to be trustful and honest, rather than misleading, or strategically vague; and (ii) whether communication is bilateral or unilateral. Taken together, these dimensions generate a natural gradient in how informative we should expect language to be about subsequent actions.

Bilateral communication under aligned incentives should generate the most informative messages, whereas unilateral communication under conflicting incentives should generate the least. Mixed cases fall in between. The three games we study map naturally onto these three regions (see Figure 1):

1. *Two-way communication with aligned incentives*, illustrated by the Minimum Effort Game in Cason et al. (2019).
2. *Two-way communication with conflicting incentives*, represented by the Inter-Group Prisoner's Dilemma in Cason et al. (2019).
3. *One-way communication with conflicting incentives*, captured by the Trust (Entry) Game in Charness and Dufwenberg (2006).

In the first setting, because the Pareto-optimum is also a Nash Equilibrium, honest communication and successful coordination go hand-in-hand, so we expect messages to be largely sincere and the mapping from words to actions to be relatively transparent. In the second setting, players have a strict monetary incentive to defect even when they talk cooperatively; cheap talk and deceptive language are natural, and communication should, in principle, be much less informative about final actions. In the third setting, communication is one-way and incentives conflict: only one player speaks and has a clear incentive to misrepresent her intentions. There is no back-and-forth or observable reaction from the receiver, so any predictive signal must come from a single short message. The next subsections briefly explain each of the games in these environments.[2]

[2]Appendix A.1 provides the exact payoff structure of each game, while appendix A.2 shows example transcripts of player conversations. For additional details of each game see Appendix B.3

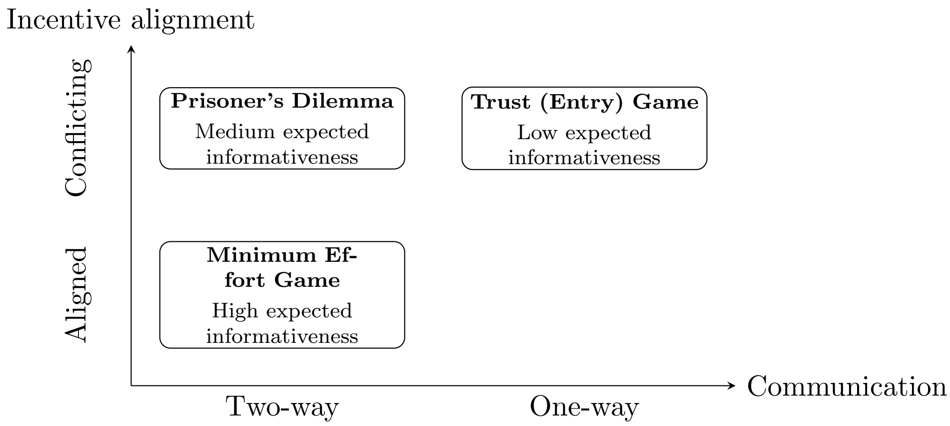


Figure 1: Strategic communication environments. The horizontal axis distinguishes bilateral from unilateral communication, and the vertical axis distinguishes aligned from conflicting incentives. The three games we study occupy different regions of this space, corresponding to different ex ante levels of language informativeness about subsequent actions.

2.2 Two-Way Communication with Aligned Incentives: Minimum Effort Game

The first environment corresponds to settings where communication facilitates coordination and incentives are fully aligned. We illustrate this with the Minimum Effort Game from Cason et al. (2019), a three-player coordination game in which each participant chooses an effort level from 1 to 7 and group earnings are determined by the minimum effort in the group. The payoff-maximizing outcome (all choosing 7) is also a Nash equilibrium, so successful coordination benefits all players (See Table A.1).

In the original experiment, groups engaged in a brief pre-play chat before making their final, anonymous choices. This chat typically involved proposing a target effort level, clarifying common interests, and confirming agreement. Because all players prefer the same cooperative outcome and no one benefits from misleading others, messages in this environment are expected to be more sincere and relatively transparent about intended actions.

Prediction Task: In our analysis, the prediction task focuses on forecasting whether a group ultimately coordinated on the socially efficient effort level —all players choose 7— (*Cooperate*) or not (*Defect*).

2.3 Two-Way Communication with Conflicting Incentives: Prisoner's Dilemma

The second environment features bilateral communication but conflicting incentives. We illustrate this with the Inter-Group Prisoner's Dilemma from Cason et al. (2019), where two groups of three participants each must choose between cooperation and defection in a one-shot Prisoner's Dilemma played at the group level. Before deciding, group members first communicate internally to form a group view, and the two groups then participate in an inter-group chat intended to establish whether mutual cooperation is feasible (See Table A.2).

The strategic structure creates clear incentives to mislead: although mutual cooperation yields the highest joint payoff, unilateral defection is individually profitable. Cooperative statements in the inter-group chat are therefore cheap talk in principle, and players may strategically offer reassurances while privately planning to defect. The combination of intra-group and inter-group communication generates rich language but also substantial scope for strategic uncertainty.

Prediction Task: The prediction task consists of forecasting the final choice of one of the two groups, but only the *other* group's internal discussion is available. In other words, forecasters observe (i) the internal chat of one group and (ii) the inter-group exchange between the two groups, but they do not observe the private, within-group conversation of the group whose action must be predicted. The objective is therefore to infer the group's intention to *Cooperate* or *Defect* solely from non-incentive compatible communication.

2.4 One-Way Communication with Conflicting Incentives: Trust (Entry) Game

The third environment features unilateral communication combined with conflicting incentives. We illustrate this with the Trust (Entry) Game from Charness and Dufwenberg (2006), a sequential interaction in which Player A first chooses whether to enter (IN) or opt out (OUT). If A enters, Player B then decides whether to take a safe payoff (DON'T ROLL) or to choose a risky option (ROLL) that can generate a higher joint surplus but exposes A to the possibility of earning zero (See Figure A.1).

Before choosing, B may send a short written message to A, but A cannot reply. The communication is therefore entirely one-way. Importantly, B has a clear monetary incentive to take the safe payoff after A enters, regardless of what is said in the message. This structure creates a setting in which messages may express reassurance, commitment, or personal reasoning, yet remain strategically ambiguous. The absence of back-and-forth interaction also makes it impossible to observe A's reactions, limiting the richness of the communicative environment.

Prediction Task: The prediction task is to forecast whether B ultimately chose ROLL (*Cooperate*) or DON'T ROLL (*Defect*). Since each prediction must rely on a single, unilateral message with no conversational context, this environment provides the thinnest language-based signal among the three settings.

3 Design and Forecasting Procedure

Our forecasting design consists of a game environment, a forecasting agent and an analysis component. The game environment comprises two inputs: The game rules and the pre-play communication chats (as described in Section 2), which are formalized into a standardized prompt. Given the prompt, forecasts are then generated by either LLMs (via API calls) or human subjects (in an online experiment). The core principle behind this design is informational symmetry: The LLM is given the same strategically relevant information as (i) the original players in past experiments at the time they made their choices and (ii) our human forecasters in Section 3.2. Finally, the analysis component records the predicted choices for evaluation.[3]

[3]The implementation described is openly accessible on GitHub.

3.1 LLM Forecasting Procedure

We use a state-of-the-art, pre-trained large language model, GPT–5, accessed via the OpenAI API, alongside earlier-generation models.[4]. Importantly, the model is *not* fine-tuned on any of the experimental datasets we study. Each prediction is obtained by querying the off-the-shelf model with a structured prompt. Temperature control had not been introduced for the OpenAI GPT–5 API at the time of the experiment, we therefore used the default API settings (including a default temperature of 1) to reflect an off-the-shelf configuration and performance quality (OpenAI (2025)).

For each game, the prompt would include: (i) a brief natural-language description of the strategic situation (roles, possible actions, and payoff-relevant consequences), (ii) the communication history available at the time of choice (multi-person chat, inter-group messages, or a single sender message, depending on the environment), and (iii) a direct question asking the model to forecast the relevant binary outcome (*Cooperate* vs. *Defect*), as defined in Sections 2.2-2.4. The prompts neither reveal the name of the original paper, nor the identity of the authors. (Prompts illustrated in Appendix B.1)

The LLMs are instructed to return its predictions in a predefined JSON format, which was introduced to allow standardization of output and facilitate automatic parsing. The output encompasses two fields: a categorical prediction (*Cooperate* vs. *Defect*) and a free-text justification.[5]

To obtain stable estimates of predictive performance, we generate multiple forecasts for each conversation. Each prompt instance corresponds to one conversation and produces a single predicted outcome. All repeated runs for a given conversation use identical inputs (same game rules and same pre-play communication text), so any variation across runs reflects only the LLM predictor's sampling stochasticity. For each of the three games, we execute the prompt 50 times (each of these executions is called a "session") with fixed parameters, thereby generating 50 forecasts for each conversation. The total number of predictions per game depends on the number of distinct conversations available in the underlying experimental dataset. Table 1 summarizes the number of conversation and the total number of LLM forecasts generated per game. Overall, we generated a total of $N = 6700$ LLM forecasts. Example raw predictions are provided in Appendix B.2.

Table 1: Summary of Experimental Groups and Forecasts by Game Type

	Minimum Effort Game	Prisoner's Dilemma	Trust (Entry) Game
Conversations	$N_{\text{MEG}} = 72$	$N_{\text{PD}} = 24$	$N_{\text{TG}} = 38$
Sessions	50	50	50
Total Forecasts	3,600	1,200	1,900

Following a standard convention in the LLM literature, we exploit the fact that API models are stateless. That is, conditional on the prompt, each call is treated as an independent draw, and requests are processed without access to prior interactions unless earlier

[4]Model 5 ("gpt-5") is used. Additional LLMs, GPT–4o ("gpt-4o") and GPT–4o mini ("gpt-4o-mini"), were also tested for between-model comparisons. For these models, we used a temperature of 0 (Section 4.3)

[5]The justification is useful for qualitative inspection, but our quantitative analysis relies only on the categorical prediction. The model is otherwise unconstrained and may use any linguistic regularities acquired during pre-training; we do not provide hand-crafted features or indicators.

messages are explicitly included (Yu et al. (2025); Bauer et al. (2023)). Hence, repeated executions of the same prompt are treated as having no carryover effects. Furthermore, since we set GPT–5's temperature at "1", some variation in the model's predictions for the same conversation is expected. However, this has negligible impact on the estimated predictive performance (See Section 4).

A potential concern is whether GPT–5 may have been exposed to the underlying experimental papers during pre-training, which could, in principle, allow trivial recognition rather than inference from the chat transcripts. Three aspects of our design help mitigate this. First, for the two games from Cason et al. (2019), neither the chat logs nor the joint distribution of messages and actions were publicly available, so high predictive accuracy at the level of individual conversations cannot stem from direct exposure to the data. Second, even though data from Trust (Entry) Game (Charness and Dufwenberg (2006)) is public, our prompts did not allow a conversation to be matched to a known observation. Third, as shown in Section 4, model performance varies across strategic settings (games) and across LLM versions in ways that are inconsistent with simple memorization: e.g. GPT–5 does not merely reproduce cooperation base rates, other models perform at or near human levels, and relative performance across games mirrors the expected informativeness of language in each strategic environment.

3.2 Human Forecasting Experiment

We ran a pre-registered human forecasting experiment[6] on Prolific on October 10, 2025 to benchmark how informative pre-play communication is for human forecasters. The design mirrors the LLM procedure as closely as possible: participants receive the same strategically relevant information as GPT–5 (see Section 3.1) and as the original players at the time of choice, and they face the same binary prediction task.

In our experiment, participants were uniformly distributed across three between-subject treatments (one per game). Each participant completed predictions for only one strategic environment, which keeps the task cognitively manageable and preserves independence across treatments. To keep the task manageable while preserving coverage of the original datasets, participants saw a subset of conversations in each treatment. In the Minimum Effort Game, each participant viewed 24 conversations drawn from a specific subtreatment of the original experiment. In the Prisoner's Dilemma, each participant viewed 12 inter-group conversations plus one randomly assigned intra-group conversation to avoid repeated exposure to the same prediction problem.[7] In the Trust (Entry) Game, participants viewed the full set of 38 sender messages. In all cases, participants predicted the relevant binary outcome (*cooperate* vs. *defect*) defined in Sections 2.2–2.4. Experimental instructions for each case can be found in Appendix B.3.

[6]The pre-registration can be found here https://aspredicted.org/38w9a9.pdf.

[7]For the Minimum Effort Game, we use the subtreatment with the highest number of defections in the original dataset, ensuring that the prediction task is non-trivial while remaining representative. For the Prisoner's Dilemma, showing repeated inter-group conversations to human participants would mechanically trivialize some forecasts, whereas zero-memory LLM queries are independent across runs and can process repeated inputs without reducing task difficulty. Appendix C.1 shows results are generally qualitatively unchanged when restricting GPT–5 to the same reduced sample used for human predictors, the only exception is that GPT–5 performs slightly better on defections (50.0%) than human forecasters (43.3%) in the Minimum Effort Game, and this difference is statistically significant according to two-sided Mann–Whitney U test.

We received a total of 201 complete responses after applying our pre-registered exclusion checks. The median completion time was 18 minutes, yielding an average payment of £12.4 per hour. Before making predictions, participants completed game-specific comprehension questions to ensure they understood the strategic structure. Following the pre-registration, participants who answered three or more comprehension questions incorrectly were removed from the study and do not appear in the final dataset.

To maintain high data quality, the study included two attention-check formats embedded within the task. First, two brief visual checks appeared at random points: a red dot was displayed for three seconds, and participants were instructed to click it while visible. This reduces the chance that subjects switch tabs or seek external help. Second, each treatment contained one "fake" conversation that initially resembled a regular transcript but later revealed itself as an attention check requiring a specific response. Robustness analyses excluding subjects according to those measures are reported in Appendix C.2. Results are qualitatively unchanged when restricting to this robust subsample.[8]

Predictions were incentivized: for each participant, one prediction was randomly selected for payment, yielding a £1 bonus if the forecast was correct. We also collected standard demographic information (age, gender, and education), which we use only for exploratory heterogeneity analyses.

4 Results

4.1 Predictive performance of GPT–5 and human forecasters

We now turn to the main empirical question of the paper: when exposed to the same pre-play conversations as incentivized human forecasters, how well does GPT–5 anticipate the actions ultimately chosen by players in each of the three strategic environments?

Table 2: Main predictive performance results

	GPT–5		**Humans**		**GPT–5 vs Humans**
Game	Accuracy (SD)	p vs. 0.5	Accuracy (SD)	p vs. 0.5	MWU p
Minimum Effort	92.4 (1.62)% *sessions = 50*	$< 0.001^{***}$	68.3 (13.5)% *n = 67*	$< 0.001^{***}$	$< 0.001^{***}$
Prisoner's Dilemma	70.3 (1.81)% *sessions = 50*	$< 0.001^{***}$	53.9 (16.4)% *n = 64*	0.073^{*}	$< 0.001^{***}$
Trust (Entry)	58.4 (2.64)% *sessions = 50*	$< 0.001^{***}$	50.8 (8.6)% *n = 71*	0.373	$< 0.001^{***}$

Notes: Each GPT–5 run and each human participant is treated as a forecaster contributing one accuracy measure (share of correct predictions within the game). Wilcoxon signed-rank tests compare accuracy to 0.5. One-sided Mann–Whitney U tests compare GPT–5 and human accuracy distributions.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

[8]Our robust sample also includes an additional filter: we exclude subjects who were not internally consistent when following simple majority voting rule in the Prisoner's Dilemma.

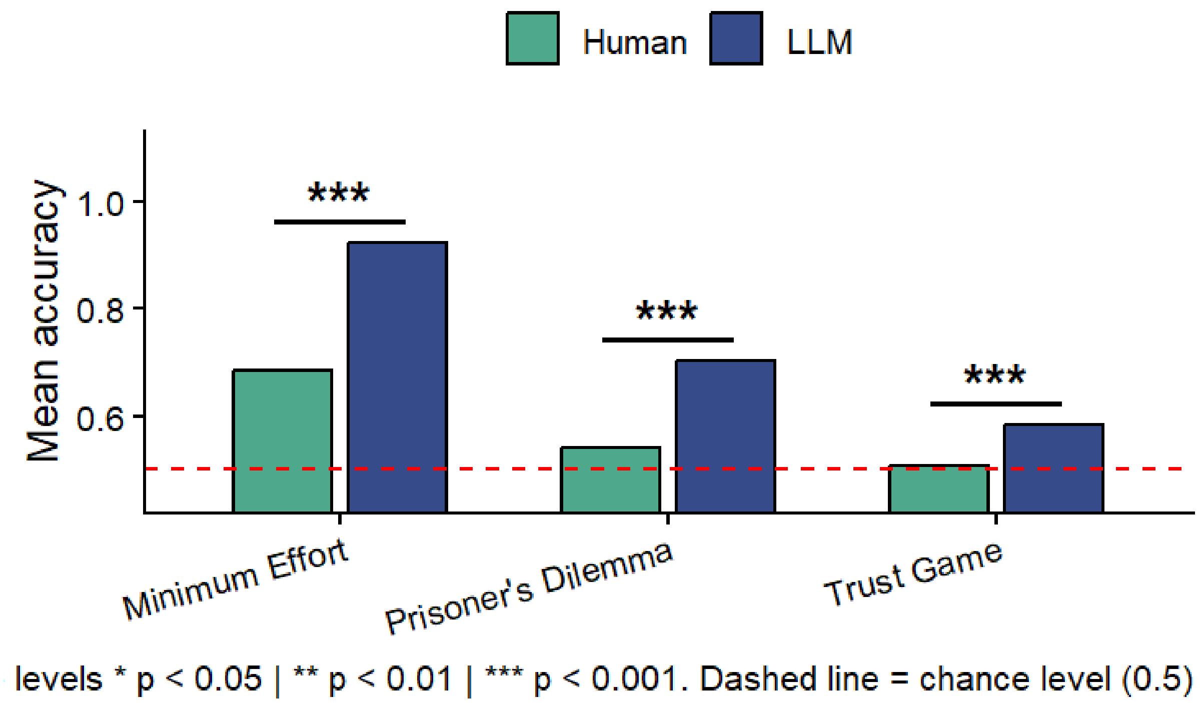


Figure 2: Mean predictive accuracy of GPT–5 and human forecasters across the three strategic environments. Bars show average accuracy; the dashed red line marks chance performance (0.5). Stars represent significance levels from one-sided Mann–Whitney U tests comparing GPT–5 to humans ($^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$). GPT-5 significantly outperforms humans in all games.

Table 2 and Figure 2 summarize the accuracy distributions for GPT–5 and human subjects. Predictive accuracy is defined at the forecaster level: one GPT–5 run refers to a set of predictions generated for all conversation groups within a given game task, which is analogous to one human participant. Each forecaster therefore contributes a single accuracy value equal to the fraction of correct predictions within the assigned game.

GPT–5 attains very high accuracy in the Minimum Effort Game, correctly forecasting outcomes in 92.4% of cases. Accuracy remains well above chance in the two environments with conflicting incentives: 70.3% in the Prisoner's Dilemma and 58.4% in the Trust (Entry) Game. In all three cases, Wilcoxon signed-rank tests reject the null of random guessing ($p < 0.001$). As Figure 2 shows, GPT–5's accuracy bars sit comfortably above the dashed 0.5 benchmark in every environment.

Human forecasters, exposed to the same transcripts, perform substantially worse. In the Minimum Effort Game, humans correctly predict 68.3% of outcomes—significantly above chance—but prediction quality deteriorates sharply in the two strategic settings with conflicting incentives (see Section 2). Accuracy averages 53.9% in the Prisoner's Dilemma and 50.8% in the Trust (Entry) Game; both of them are statistically indistinguishable from chance at the 0.05 level. Mann–Whitney U tests confirm that GPT–5 significantly outperforms human forecasters in all three environments (all $p < 0.001$). The performance gaps are large: over 20 percentage points in the Minimum Effort Game and roughly 17–20 points in the two environments with conflicting incentives.

Result 1: GPT–5 predicts strategic behavior well above chance in all three strategic environments.

Result 2: GPT–5 significantly outperforms human forecasters in all three strategic environments.

Result 3: Human forecasters only predict above chance in the Minimum Effort Game.

Taken together, these results show that GPT–5 reliably performs well-above random and markedly better than human forecasters across all strategic settings. The next subsection examines the structure of GPT–5's predictive behavior across environments and explores how performance relates to the nature of the underlying communication settings.

4.2 Patterns in Predictive Behavior

We now examine how predictive performance varies across strategic environments. First, we describe regularities in GPT–5's forecasts once we take into account the structural differences between the strategic environments described in Section 2. These environments differ in whether communication is bilateral or unilateral and in whether incentives make honest messages more or less likely, generating a conceptual gradient in how informative pre-play communication is expected to be about subsequent actions.

Table 2 and Figure 2 already showed a clear ordering in predictive accuracy across the three settings. GPT–5 performs best in the Minimum Effort Game (92.4%), less well in the Prisoner's Dilemma (70.3%), and least well in the Trust (Entry) Game (58.4%). Human forecasters display exactly the same monotonic pattern, although at substantially lower accuracy levels (68.3%, 53.9%, 50.8%). This descriptive ordering aligns naturally with the taxonomy introduced in Section 2.1: environments in which communication is expected to be more informative are precisely those in which both GPT–5 and humans achieve higher predictive performance.

Result 4: GPT–5's accuracy increases monotonically from the environment with the highest expected informativeness (*two-way communication with aligned incentives*) to the environment with the lowest (*one-way communication with conflicting incentives*); human forecasters display the same pattern.

A natural benchmark for assessing whether GPT–5's predictions exploit information in communication is the empirical cooperation frequency in each game. Table 3 reports the cooperation base-rates—90.3% in the Minimum Effort Game, 66.67% in the Prisoner's Dilemma, and 65.79% in the Trust (Entry) Game—alongside human and GPT–5 conditional accuracy on cooperative and defective outcomes. Note that a forecaster that always predicted "cooperate" would achieve accuracy equal to the cooperation base-rate. Comparing GPT–5's performance to this benchmark therefore provides a direct test of whether its accuracy could be explained by unconditional forecasting, irrespective of the informativeness of communication.

To evaluate this formally, we use Fisher's Exact test to examine whether GPT–5's predictions are statistically associated with the realized cooperation or defection outcomes. For each game, Table 3 reports GPT–5's conditional accuracy given cooperation and given

Table 3: Conditional Accuracy and Prediction–Outcome Association (GPT–5 vs Humans)

Game	Cooperation LLM	Cooperation Human	Cooperation MWU	Defection LLM	Defection Human	Defection MWU	Fisher p-value	Base Rate (%)
Minimum Effort	99.3	73.3	$< 0.001^{***}$	28.6	43.3	0.6637	$< 0.001^{***}$	90.3
Prisoner's Dilemma	98.9	69.7	$< 0.001^{***}$	13.2	22.7	0.9046	$< 0.001^{***}$	66.66
Trust (Entry)	72.4	60	$< 0.001^{***}$	31.4	33.2	0.8184	0.088^{*}	65.79

Notes: Conditional accuracy = share of correct predictions conditional on the true outcome. Two-sided MWU tests compare LLM vs human accuracy. Fisher exact tests evaluate whether predicted and realized actions are associated (LLM only). Base rate refers to empirical cooperation frequencies in the original games.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

defection together with the corresponding Fisher Exact p-values. If GPT–5 were relying on an unconditional forecasting rule (such as always predicting cooperation or applying any other rule that does not condition on the content of the messages) its predicted frequency of cooperation would be the same across conversations that ultimately end in cooperation and those that end in defection. In that case, predicted actions and realized outcomes would be statistically independent, and Fisher's Exact test would not reject independence. Instead, we find strong evidence of association in both the Minimum Effort Game and the Prisoner's Dilemma (both $p < 0.001$), indicating that GPT–5 systematically varies its forecasts across conversations in a way that tracks subsequent actions rather than relying on an unconditional rule. In the Trust (Entry) Game, Fisher's Exact test plays a less central role because the key diagnostic comes from the comparison between accuracy and the cooperation base rate. Cooperation occurs in 65.79% of cases, yet GPT–5's accuracy is only 58.4%—a gap of 7.4 percentage points (more than three standard deviations given SD = 2.64%)—which rules out an unconditional "always cooperate" forecasting strategy by construction.

Result 5: GPT–5's predictive performance reflects systematic sensitivity to conversational content rather than replication of unconditional cooperation frequencies across all three strategic environments.

A complementary perspective comes from examining GPT–5's predictive performance conditional on the true outcome. Table 3 and Figure 3 show that GPT–5's high overall accuracy is driven primarily by its ability to identify cooperative outcomes with remarkable precision. In the Minimum Effort Game and the Prisoner's Dilemma, GPT–5 correctly forecasts cooperation in 99.3% and 98.9% of cases respectively. These levels far above those of human forecasters (73.3% and 69.7%, both $p < 0.001$). Even in the Trust (Entry) Game, where communication is unilateral and incentives conflict, GPT–5 attains 72.4% accuracy on cooperative outcomes, again significantly outperforming humans ($p < 0.001$).

In contrast, performance on defections is markedly lower. GPT–5 correctly predicts only 28.6% of defections in the Minimum Effort Game, 13.2% in the Prisoner's Dilemma, and 31.4% in the Trust (Entry) Game. Human forecasters perform slightly better on defections (43.3%, 22.7%, and 33.2% respectively) but these differences are not statistically significant according to two-sided Mann–Whitney U tests (all $p > 0.66$). The pattern that emerges is therefore asymmetric: GPT–5 excels at detecting cooperative intent but performs substantially less well in distinguishing conversations that end in defection.

The combination of these results helps clarify the source of GPT–5's predictive advantage. Its strong performance does not arise because it predicts cooperation indiscriminately—indeed, the Fisher Exact tests and the accuracy–base-rate comparison rule out such a heuristic. Rather, GPT–5 appears highly sensitive to linguistic cues associated with cooperative behavior, whereas conversational signals preceding defection are either weaker, more heterogeneous, or strategically noisier. Overall, the conditional results indicate that GPT–5 extracts reliable cues associated with cooperative behavior, whereas signals associated with defection may appear weaker or noisier, making defection harder to predict.

Result 6: GPT–5 predicts cooperative outcomes with substantially higher accuracy than human forecasters in all three environments.

Result 7: GPT–5's accuracy on defections is comparatively low and not significantly different from that of human forecasters.

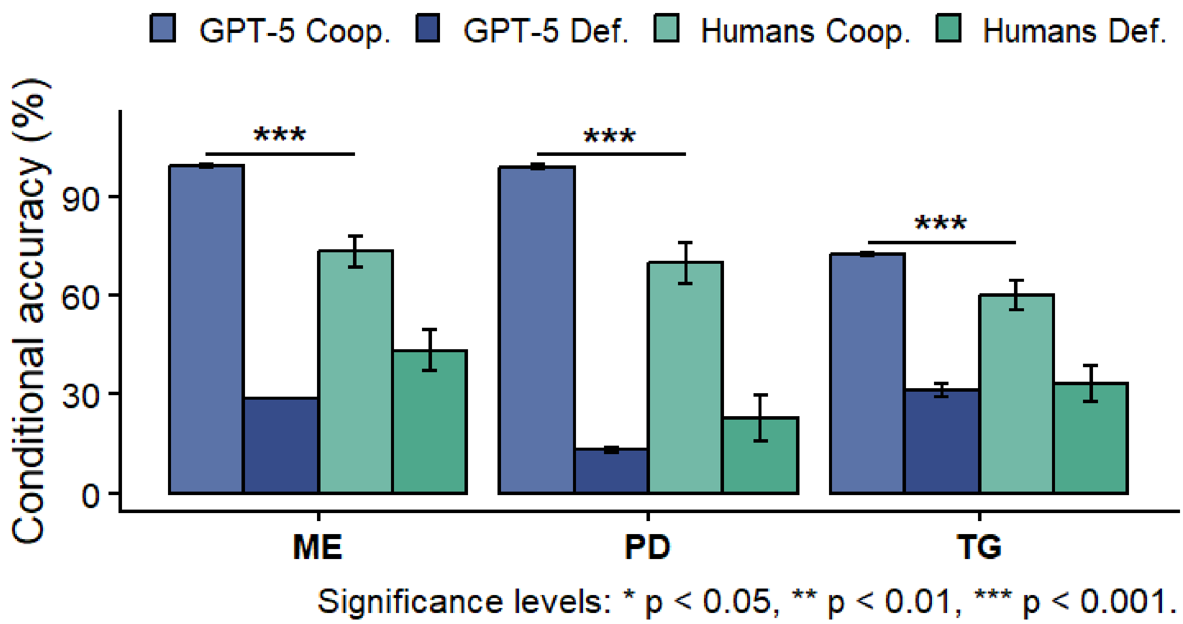


Figure 3: Conditional accuracy of GPT-5 and human forecasters based on Cooperation or Defection outcomes. Error bars are confidence intervals at the 95% level. Stars indicate Mann–Whitney U tests comparing GPT-5 vs. Humans for the same outcome (Cooperate or Defect). No stars indicate non-significant differences.

4.3 Model Comparisons

The main purpose of this subsection is to assess whether predictive performance is specific to GPT–5 or reflects a broader capability of contemporary LLMs. Table 4 reports forecaster-level accuracy for GPT–5, GPT–4o, GPT–4o mini, and human subjects, while Figure 4 visualizes these differences across the three environments. A remarkable feature of the results is that even the weakest-performing model in each game (GPT–4o mini in the Minimum Effort and Trust (Entry) Games, and GPT–4o in the Prisoner's Dilemma) performs at least as well as human forecasters, and significantly better in two of the three environments (MWU $p = 0.016$ in the Minimum Effort Game and $p < 0.001$ in the Prisoner's Dilemma). In the Trust (Entry) Game, the weakest model performs comparably to humans ($p = 0.337$). Thus, the ability to forecast human strategic behavior is not confined to the most advanced model: it is shared across different LLM models and persists even for smaller model[9].

Table 4: Predictive performance across models and humans

Game	Accuracy (SD)				MWU p-values		
	GPT–5	GPT–4o	GPT–4o mini	Humans	5 vs 4o	5 vs mini	WPM vs Humans
Minimum Effort Game Mean (SD) %	92.4 (1.62) *sessions=50*	91.6 (0.33) *sessions = 50*	91.5 (0.67) *sessions = 50*	68.3 (13.5) *n = 67*	$< 0.001^{***}$	$< 0.001^{***}$	$< 0.001^{***}$
Prisoner's Dilemma Mean (SD) %	70.3 (1.81) *sessions = 50*	59.4 (4.97) *sessions = 50*	70.8 (0.00) *sessions = 50*	53.9 (16.4) *n = 71*	$< 0.001^{***}$	0.026^{**}	0.016^{**}
Trust (Entry) Game Mean (SD) %	58.4 (2.64) *sessions = 50*	53.2 (2.01) *sessions = 50*	51.1 (2.77) *sessions = 50*	50.8 (8.6) *n = 64*	$< 0.001^{***}$	$< 0.001^{***}$	0.337

Notes: Accuracy is defined at the forecaster level. Mann–Whitney U tests compare the distributions of forecaster-level accuracies for the specified pairs: GPT–5 vs GPT–4o, GPT–5 vs GPT–4o mini, and the worst-performing model (WPF) vs humans (within each game). Two-sided for model comparisons, one sided for model vs humans.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Within this broader pattern, GPT–5 consistently attains the highest predictive accuracy. In the Minimum Effort Game, all three models achieve accuracy above 91%, with GPT–5 modestly but significantly outperforming the others. Differences become more pronounced in the Prisoner's Dilemma, where GPT–5 reaches 70.3% accuracy compared to 59.4% for GPT–4o and 70.8% for GPT–4o mini, and especially in the Trust (Entry) Game, where GPT–5 achieves 58.4% accuracy versus 53.2% and 51.1% for the weaker models. These patterns indicate that accuracy differences across model families widen as the predictive environment becomes more challenging, even though all models remain at or above human performance.

Taken together, the results in Table 4 and Figure 4 show that GPT–5 provides the strongest forecasts, but that the broader phenomenon is not model-specific: across tasks and environments, LLMs of varying capacities reliably match or exceed human predictive performance.

Result 8: In all three environments, the weakest-performing model's accuracy is either significantly higher than or statistically indistinguishable from human accuracy.

Result 9: GPT–5 achieves higher accuracy than GPT–4o and GPT–4o mini in all three environments.

[9]While parameter counts are not explicitly published for comparison between GPT–5 and GPT–4, OpenAI has publicized GPT–4o mini as a small, lightweight model OpenAI (2024).

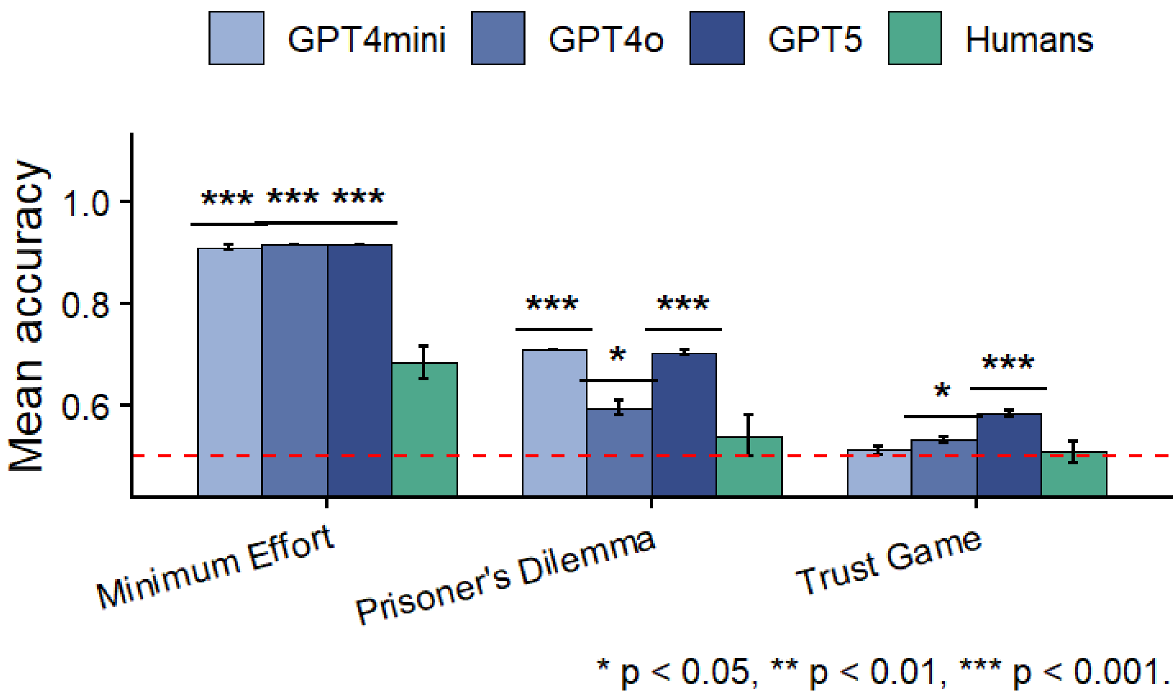


Figure 4: Predictive performance across models and humans and all games. Error bars are confidence intervals at 95% level. Stars indicate one-sided Mann–Whitney U tests comparing each model's accuracy distribution to the human accuracy distribution within the same game. The dashed line represents chance level, identical to 50% accuracy.

4.4 Individual Heterogeneity: "Superforecasters"

Given the performance of GPT–5 and humans, a natural question is whether some individuals can consistently outperform GPT–5 and if so, what are their characteristics. The findings below are exploratory and were not pre-registered.

We define a "superforecaster" (Tetlock and Gardner, 2016) as an individual whose forecast accuracy *strictly outperforms* GPT–5 in a given environment. More specifically, we evaluate each individual's accuracy against GPT–5's mean accuracy. The density plot in Figure 5 shows the smoothed-out distribution of human performance against GPT–5 benchmark (vertical line). The benchmark lies above the bulk of the human distribution, indicating it outperforms most participants. However, heterogeneity in human performance is evident, GPT–5 does not dominate all individuals in every game, a meaningful minority of individuals still exceed GPT–5's mean performance. In the Minimum Effort Game, no individual outperforms GPT–5; in the Prisoner's Dilemma, and in the Trust (Entry) Game, 11 and 8 individuals outperform the LLM respectively.

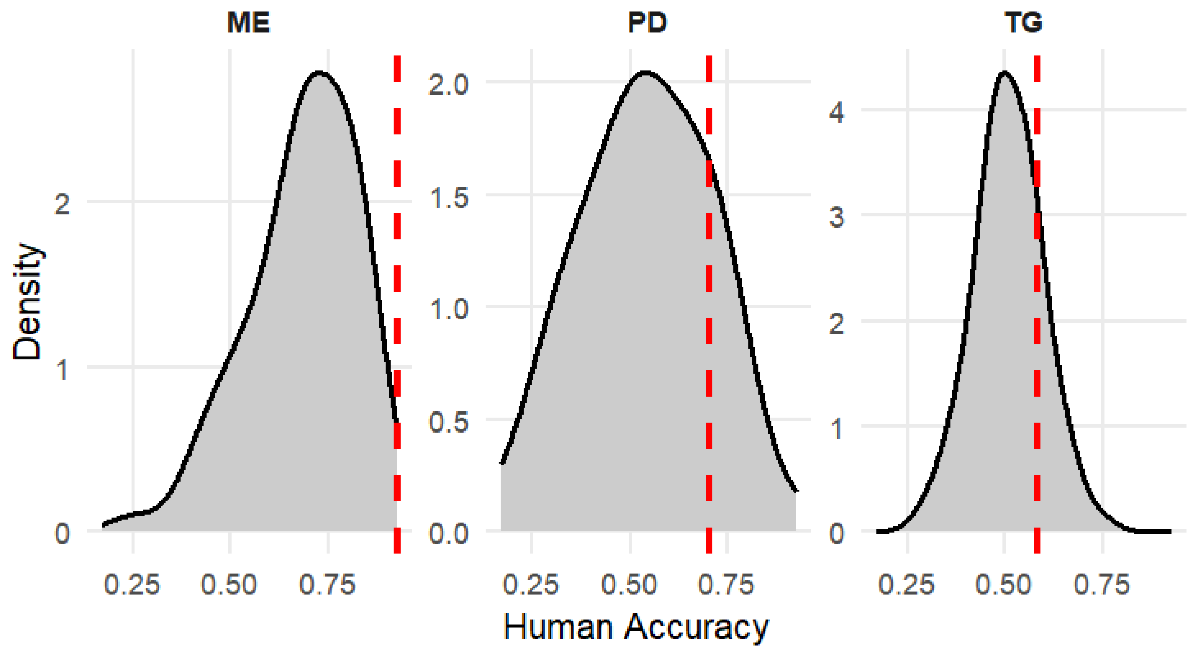


Figure 5: Human forecasting performance, density plots show human performances, red line shows GPT–5 mean

Once again, following the strategic environments described in Section 2, we observe slightly more "superforecasters" in two-way communication with conflicting incentives. This pattern suggests that comparatively, more humans might be able to outperform GPT–5 when strategic environments contain conflicting incentives.

In Table 5, we reported the logistic regression results of being a "superforecaster" on gender, age and education, estimated separately by games (i.e. Prisoner's Dilemma and Trust (Entry) Game respectively). Education is significantly correlated with the "superforecaster" status in Prisoner's Dilemma, with higher education associated with higher likelihood of being classified as a "superforecaster". Gender is weakly significant and positively correlated to being a "superforecaster" in the Trust (Entry) Game. These results show that different demographic characteristics are indicative of exceptional performance across games. However, given the small sample size, and the limited demographic information available, this evidence remain suggestive.

Table 5: Demographic correlates of superforecaster status

	Game	
Variable	**Prisoner's Dilemma**	**Trust (Entry) Game**
Female	-1.625	1.880^{*}
	(1.001)	(1.122)
Age	0.303	0.00111
	(0.366)	(0.331)
Education	1.825^{**}	0.195
	(0.777)	(0.327)
Constant	-9.557^{**}	-4.084^{**}
	(3.813)	(1.880)
Observations	61	70
Pseudo-R^2	0.289	0.092

Notes: Logistic regressions estimate the association between demographic characteristics and superforecaster status, separately by game. Robust standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

5 Conclusion

We study whether modern large language models can anticipate strategic behavior from human-to-human communication, and how their performance compares to that of humans facing the same prediction task. Across three canonical environments, we find that GPT–5 outperforms incentivized human forecasters when predicting future actions from pre-play transcripts. The advantage is especially pronounced when incentives are aligned and communication is bilateral, suggesting that the model is able to detect subtle linguistic signals of cooperative intent that humans either overlook or discount. By contrast, neither humans nor the model perform well when incentives conflict and messages are likely to be strategically misleading.

These results indicate that pre-play communication often contains more information about intentions than humans are able to extract in real time. Large language models, trained on large data, appear to map conversational patterns to likely outcomes more effectively than humans. This sheds light on the cognitive limits of human forecasting within strategic contexts and raises new questions about how beliefs are formed from communication.

This ability to predict behavior from deliberation and negotiation transcripts has implications for real-world settings in which decisions follow communication: regulatory committees, collective bargaining, diplomacy, or online market interactions where promises precede actions. Our controlled environment provides a benchmark before moving toward naturally occurring communication in more complex institutional settings.

An important next step for future research could be to move from forecasting performance to specific linguistic mechanisms: That is, which linguistic features convey strategic intent, and why are humans insensitive to them? While this paper did not explore how LLM-text classification (Gentzkow et al., 2019; Celebi and Penczynski, 2025) and prediction interact, by identifying the structure of persuasive and diagnostic language, future work can shed light on what communication reveals, when it misleads, and how humans and algorithms differ in the interpretation of strategic messages.

References

Agranov, M. (2025): "Communication in Games," in *Handbook of Experimental Economics Methods*, ed. by E. Snowberg and L. Yariv, North-Holland/Elsevier, forthcoming. Cited on pages 1 and 2.

Agranov, M. and C. Tergiman (2019): "Communication in bargaining games with unanimity," *Experimental Economics*, 22, 350–368. Cited on page 3.

Akata, E., L. Schulz, J. Coda-Forno, S. J. Oh, M. Bethge, and E. Schulz (2025): "Playing repeated games with large language models," *Nature Human Behaviour*, 1–11. Cited on page 3.

Alempaki, D., V. Burdea, and D. Read (2024): "Deceptive communication: direct Lies vs. ignorance, selective truth and silence," . Cited on page 3.

Apel, R., I. Erev, R. Reichart, and M. Tennenholtz (2022): "Predicting decisions in language based persuasion games," *Journal of Artificial Intelligence Research*, 73, 1025–1091. Cited on page 3.

Backus, M., T. Blake, and S. Tadelis (2019): "On the empirical content of cheap-talk signaling: An application to bargaining," *Journal of Political Economy*, 127, 1599–1628. Cited on page 3.

Bauer, K., L. Liebich, O. Hinz, and M. Kosfeld (2023): "Decoding gpt's hidden 'rationality'of cooperation," . Cited on page 8.

Ben-Ner, A. and L. Putterman (2009): "Trust, communication and contracts: An experiment," *Journal of Economic Behavior & Organization*, 70, 106–121. Cited on page 3.

Ben-Ner, A., L. Putterman, and T. Ren (2011): "Lavish returns on cheap talk: Two-way communication in trust games," *The Journal of Socio-Economics*, 40, 1–13. Cited on page 3.

Blume, A. and A. Ortmann (2007): "The effects of costless pre-play communication: Experimental evidence from games with Pareto-ranked equilibria," *Journal of Economic theory*, 132, 274–290. Cited on page 3.

Bond, C. F. and B. M. DePaulo (2006): "Accuracy of Deception Judgments," *Personality and Social Psychology Review*, 10, 214–234. Cited on page 1.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, et al. (2020): "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, 33, 1877–1901. Cited on page 1.

Burton, A. and M. Sefton (2004): "Risk, pre-play communication and equilibrium," *Games and economic behavior*, 46, 23–40. Cited on page 3.

Cason, T. N., S.-H. P. Lau, and V.-L. Mui (2019): "Prior interaction, identity, and cooperation in the Inter-group Prisoner's Dilemma," *Journal of Economic Behavior & Organization*, 166, 613–629. Cited on pages 1, 4, 5, 8, and 23.

Celebi, C. and S. Penczynski (2025): "Using Large Language Models for Text Classification in Experimental Economics," Tech. rep., School of Economics, University of East Anglia, Norwich, UK. Cited on page 17.

Charness, G. and M. Dufwenberg (2006): "Promises and partnership," *Econometrica*, 74, 1579–1601. Cited on pages 1, 3, 4, 6, 8, and 24.

Charness, G., F. Feri, M. A. Meléndez-Jiménez, and M. Sutter (2023): "An experimental study on the effects of communication, credibility, and clustering in network games," *Review of Economics and Statistics*, 105, 1530–1543. Cited on page 3.

Cooper, D. J. and J. H. Kagel (2023): "Using team discussions to understand behavior in indefinitely repeated prisoner's dilemma games," *American Economic Journal: Microeconomics*, 15, 114–145. Cited on page 3.

Cooper, R., D. V. DeJong, R. Forsythe, and T. W. Ross (1989): "Communication in the battle of the sexes game: some experimental results," *The RAND Journal of Economics*, 568–587. Cited on page 1.

Costa-Gomes, M. A. and G. Weizsäcker (2008): "Stated Beliefs and Play in Normal-Form Games," *Review of Economic Studies*, 75, 729–762. Cited on page 1.

Crawford, V. P. and J. Sobel (1982): "Strategic Information Transmission," *Econometrica*, 50, 1431–1451. Cited on page 1.

Duan, J., R. Zhang, J. Diffenderfer, B. Kailkhura, L. Sun, et al. (2024): "GTBench: Uncovering the Strategic Reasoning Limitations of LLMs via Game-Theoretic Evaluations," *arXiv preprint arXiv:2402.12348*. Cited on page 3.

Duffy, J. and N. Feltovich (2002): "Do actions speak louder than words? An experimental comparison of observation and cheap talk," *Games and Economic Behavior*, 39, 1–27. Cited on page 3.

——— (2006): "Words, deeds, and lies: Strategic behaviour in games with multiple signals," *The Review of Economic Studies*, 73, 669–688. Cited on page 3.

Ellingsen, T. and R. Östling (2010): "When does communication improve coordination?" *American Economic Review*, 100, 1695–1724. Cited on page 3.

Erva Ergun, Z. and E. Sefer (2025): "Financial Statement Fraud Detection via Large Language Models," *Intelligent Systems in Accounting, Finance and Management*, 32, e70021. Cited on page 3.

Farrell, J. and M. Rabin (1996): "Cheap Talk," *Journal of Economic Perspectives*, 10, 103–118. Cited on page 1.

Feltovich, N. and J. Swierzbinski (2011): "The role of strategic uncertainty in games: An experimental study of cheap talk and contracts in the nash demand game," *European Economic Review*, 55, 554–574. Cited on page 3.

Fonseca, M. A. and H.-T. Normann (2012): "Explicit vs. tacit collusion—The impact of communication in oligopoly experiments," *European economic review*, 56, 1759–1772. Cited on page 1.

Gentzkow, M., B. Kelly, and M. Taddy (2019): "Text as data," *Journal of Economic Literature*, 57, 535–574. Cited on page 17.

Gneezy, U. (2005): "Deception: The Role of Consequences," *American Economic Review*, 95, 384–394. Cited on page 1.

Greiner, B., M. Caravella, and A. E. Roth (2014): "Is avatar-to-avatar communication as effective as face-to-face communication? An Ultimatum Game experiment in First and Second Life," *Journal of Economic Behavior & Organization*, 108, 374–382. Cited on page 3.

Guo, S., H. Bu, H. Wang, Y. Ren, D. Sui, Y. Shang, and S. Lu (2024): "Economics arena for large language models," *arXiv preprint arXiv:2401.01735*. Cited on page 3.

Hazra, S. and B. P. Majumder (2024): "To tell the truth: Language of deception and language models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8506–8520. Cited on page 3.

Horton, J. J. (2023): "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" NBER Working Paper 31122, National Bureau of Economic Research. Cited on page 3.

Kawagoe, T. and H. Takizawa (2009): "Equilibrium refinement vs. level-k analysis: An experimental study of cheap-talk games with private information," *Games and Economic Behavior*, 66, 238–255. Cited on page 3.

Kirkos, E., G. Boskou, E. Chatzipetrou, E. Tiakas, and C. Spathis (2024): "Exploring the boundaries of financial statement fraud detection with large language models," *Available at SSRN 4842962*. Cited on page 3.

Kretschmar, K., W. Laurito, S. Maiya, and S. Marks (2025): "Liars' Bench: Evaluating Lie Detectors for Language Models," *arXiv preprint arXiv:2511.16035*. Cited on page 3.

Li, X., Ö. Özer, and U. Subramanian (2022): "Are we strategically naïve or guided by trust and trustworthiness in cheap-talk communication?" *Management Science*, 68, 376–398. Cited on page 3.

Loconte, R., R. Russo, P. Capuozzo, P. Pietrini, and G. Sartori (2023): "Verbal lie detection using large language models," *Scientific reports*, 13, 22849. Cited on page 3.

Luo, X., A. Rechardt, G. Sun, K. K. Nejad, F. Yáñez, B. Yilmaz, and B. C. Love (2025): "Large language models surpass human experts in predicting neuroscience results," *Nature Human Behaviour*, 9, 305–315. Cited on page 4.

Meloso, D., S. Nunnari, and M. Ottaviani (2023): "Looking into crystal balls: a laboratory experiment on reputational cheap talk," *Management Science*, 69, 5112–5127. Cited on page 3.

Moreno, D. and J. Wooders (1998): "An experimental study of communication and coordination in noncooperative games," *Games and Economic Behavior*, 24, 47–76. Cited on page 3.

OpenAI (2024): "GPT-4o mini: advancing cost-efficient intelligence," Accessed: 2025-12-20. Cited on page 14.

OpenAI (2025): "Latestmodel parameter compatibility," https://platform.openai.com/docs/guides/latest-model, accessed: 2025-11-15. Cited on page 7.

Palfrey, T., H. Rosenthal, and N. Roy (2017): "How cheap talk enhances efficiency in threshold public goods games," *Games and Economic Behavior*, 101, 234–259. Cited on page 3.

Proto, E., D. Sgroi, and M. Nazneen (2017): "The effect of positive mood on cooperation in repeated interaction," Tech. rep., IZA Discussion Papers. Cited on page 3.

Rogiers, A., S. Noels, M. Buyl, and T. De Bie (2024): "Persuasion with large language models: a survey," *arXiv preprint arXiv:2411.06837*. Cited on page 3.

Sánchez-Pagés, S. and M. Vorsatz (2009): "Enjoy the silence: an experiment on truth-telling," *Experimental Economics*, 12, 220–241. Cited on page 3.

Shahriar, S. (2025): "Linguistic Deception Detection–Models, Domains, Behaviors, Stylistic Patterns to Large Language Models (LLMs)," Ph.D. thesis. Cited on page 3.

Sibly, H. and J. Tisdell (2018): "Cooperation and turn taking in finitely-repeated prisoners' dilemmas: An experimental analysis," *Journal of Economic Psychology*, 64, 49–56. Cited on page 3.

Sutter, M. and C. Strassmair (2009): "Communication, cooperation and collusion in team tournaments—an experimental study," *Games and Economic Behavior*, 66, 506–525. Cited on page 3.

Tetlock, P. E. and D. Gardner (2016): *Superforecasting: The art and science of prediction*, Random House. Cited on page 15.

Vanberg, C. (2008): "Why Do People Keep Their Promises? An Experimental Test of Two Explanations," *Econometrica*, 76, 1467–1480. Cited on page 1.

Yu, L., J. Lin, and J. Li (2025): "Stateful large language model serving with pensieve," in *Proceedings of the Twentieth European Conference on Computer Systems*, 144–158. Cited on page 8.

Zultan, R. (2012): "Strategic and social pre-play communication in the ultimatum game," *Journal of Economic Psychology*, 33, 425–434. Cited on page 3.

Appendix

A Experimental Environments

A.1 Payoff and Structure

- Payoff Structure of the Minimum-Effort Game: Table A.1
- Payoff Structure of the Prisoner's Dilemma: Table A.2
- Game and payoff structure of the Trust (Entry) Game: Figure A.1

A.1.1 Payoff Structure of the Minimum-Effort Game

	Smallest number in the group						
Own number	7	6	5	4	3	2	1
7	19.50	16.50	13.50	10.50	7.50	4.50	1.50
6	-	18.00	15.00	12.00	9.00	6.00	3.00
5	-	-	16.50	13.50	10.50	7.50	4.50
4	-	-	-	15.00	12.00	9.00	6.00
3	-	-	-	-	13.50	10.50	7.50
2	-	-	-	-	-	12.00	9.00
1	-	-	-	-	-	-	10.50

Table A.1: Payoffs of the minimum-effort game

Notes: Payoffs in Cason et al. (2019), the row indicates a player i's own choice, the column the minimum in the group (including player i).

A.1.2 Payoff Structure of the Prisoner's Dilemma

	Cooperate	Defect
Cooperate	$132, 132$	$28, 162$
Defect	$162, 28$	$54, 54$

Table A.2: Prisoner's dilemma game of Cason et al. (2019).

A.1.3 Payoff Structure of the Trust (Entry) Game

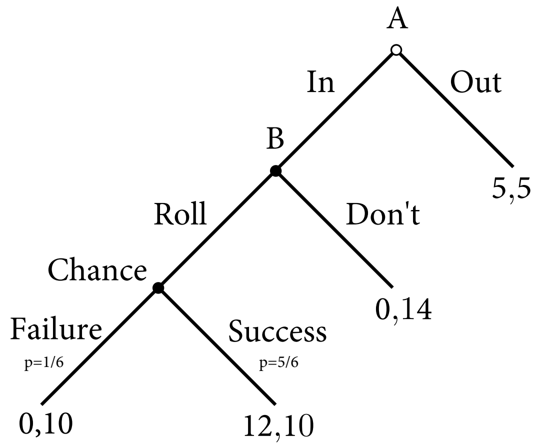


Figure A.1: Trust game with hidden action of Charness and Dufwenberg (2006).

A.2 Example Conversations

- Example Conversation of the Minimum Effort Game: Figure A.2
- Example Conversation of the Prisoner's Dilemma (Intra-Group Chat): Figure A.3
- Example Conversation of the Prisoner's Dilemma (Inter-Group Chat): Figure A.4
- Example Conversation of the Trust Entry Game: Figure A.5

Minimum Effort Game: Example Chat Log

Proposals:
Player 8: **7**
Player 10: **7**
Player 12: **5**

Chat Log:

> Player 10: "hey guys, why don't we all choose 7?"
> Player 8: "Please choose 7!"
> Player 12: "ok."
> Player 10: "to maximize our payoff!! :)"
> Player 8: "So, we will all choose 7!"
> Player 10: "Deal!"
> Player 10: "GS 2?"
> Player 12: "DEAL!!!!!!"
> Player 10: "Excellent."
> Player 8: "Deal!"

Figure A.2: Example of pre-play communication in the Minimum Effort Game.

Inter-Group Prisoner's Dilemma: Intra-Group Chat (Example)

Player C: "hello"
Player C: "do you think we should risk?"
Player A: "hi all"
Player B: "finally some difference"
Player C: "i think as long as they don't trick us"
Player C: "we should stick with M"
Player C: "BUT if they trick us, then we'll get 28 :("
Player B: "that's the problem"
Player C: "should we trick them?"
Player B: "but this is a game"
Player A: "i dunno agree, in either scenario, choosing M is always not appropriate"
Player C: "then we switch to J :P"
Player C: "TRUE"
Player B: "let's say we all choose M in the second chat"
Player A: "yes.. then we switch to J"

Figure A.3: Example intra-group communication in the Inter-Group Prisoner's Dilemma.

Inter-Group Prisoner's Dilemma: Inter-Group Chat (Example)

Player C: "hello"
Player 3: "hey"
Player 3: "hi"
Player B: "hi"
Player A: "hey"
Player C: "maybe we should all choose M?"
Player 3: "=)"
Player 1: "alright, you guys all know the proposition"
Player 3: "both choose M is the win–win situation"
Player C: "it seems to be better for both of us"
Player A: "let's stick with M"
Player 3: "yes"
Player B: "I think so"

Figure A.4: Example inter-group communication in the Inter-Group Prisoner's Dilemma.

Trust (Entry) Game: Example Sender Messages

Excerpt 1:

> "Hey. OK I think that the best way for both of us to make a profit is for you to choose IN and for me to roll. That way we both make some money. There's no point in me not rolling because that would give you and me less profit. So I'm a roller if you're in :)"

Excerpt 2:

> "I have to do laundry tonight and I really don't want to do it! But I don't have any clean underwear left and I don't want to go commando tomorrow. We'll see what I decide tonight. This man acts funny doesn't he? But he seems cool, he's quite a character. All this mystery is kinda cool."

Figure A.5: Example one-way communication in the Trust (Entry) Game.

B Forecasting Design

B.1 LLM Prompts

B.1.1 Prompt for the Minimum Effort Game

SYSTEM MESSAGE

You are an assistant trained to analyze and predict whether a group of university students playing a **minimum effort game** will successfully coordinate on the highest number (7) or fail to coordinate (at least one player chooses less than 7). This game was conducted as an experiment in 2018.

Note: Always use the financial incentives and the chat dynamics to guide your prediction. Focus on identifying signs of group consensus or disagreement.

Game Context

- Players are university students participating in a controlled experiment to study decision-making and coordination.
- The **minimum effort game** involves a group of 3 players, each independently selecting an integer between **1 and 7**.
- The group's payoff is determined by the lowest number chosen (the group's "minimum effort").
- If all players choose 7, the group achieves the highest possible payoff.
- If at least one player chooses less than 7, the group's payoff is reduced.
- However, if one person deviates the relative monetary amount will be higher for the one deviating. This is crucial for those who may misunderstand the game as a competition, even if it hurts their own monetary payoff.
- Each player in the game is given a name, consisting of the term "Gamma Square" (GS) followed by a number.

Financial Incentives

- Players receive a monetary reward based on the group's minimum effort:
- Choosing 7 yields the highest individual and group payoff if all players coordinate.
- Choosing a lower number results in a reduced payoff for the group, with no individual advantage.
- Players may fail to coordinate if they misunderstand the game, feel insecure about others' choices, or wish to undermine the group intentionally.

Your Task

Based on the proposals (non-binding initial choices) and the chat logs, predict whether the group will coordinate successfully (all players choose 7) or fail to coordinate (at least one player chooses less than 7).

Steps to Follow

1. *Analyze Proposals*: Use the proposals as an initial signal of each player's intentions.
2. *Predict Outcome*: Based on the proposals and chat content, determine whether the group will:
   - Coordinate (all players choose 7).
   - Fail to Coordinate (at least one player chooses less than 7).

Output Format

Your output should follow EVERY of the steps above and give a lot of importance to misunderstanding. Then predict WHAT NUMBER EACH player will choose. E.g. You must specify: Player 8 will play 7, Player 10 will play 8 etc. In the end, conclude with "Coordinate" or "Fail to Coordinate" depending on whether all players are predicted to choose 7 or not.

Your output MUST be a single, valid JSON object and nothing else. Do not include any text before or after the JSON object.

JSON Output Structure

You must generate a JSON object with the following exact structure:

```
{
  "player_predictions": [
    {
      "player_id": "string",
      "predicted_choice": integer,
      "prediction_context": "A string explaining the prediction.
          CRITICAL: If the prediction is based on the player's
          proposal, this string MUST include the word 'proposal'.
          If it is based on their chat messages, it MUST include
          the word 'chat'."
    }
  ],
  "conclusion": {
    "outcome": "A string, either 'Coordinate' or 'Fail to
        Coordinate'.",
    "explanation": "A string providing a final summary and
        explanation for the overall outcome."
  }
}
```

USER MESSAGE

Important: Please bear in mind that proposals come before chatting.

Please analyze the following minimum effort game session and predict the outcome:

Initial Proposals (Non-binding)

`{PROPOSALS_DATA}`

Chat Logs

`{CHAT_LOGS}`

Based on this information, please provide your analysis following the framework outlined in the system message.

B.1.2 Prompt for the Prisoner's Dilemma

SYSTEM MESSAGE

You are an assistant trained to analyze and predict whether a team of three university students each playing an **Inter-Group Prisoner's Dilemma** will cooperate (at least two players choose "cooperate") or defect (at least two players choose "defect"). This game was conducted as an experiment in 2018.

Game Context

- A group of university students is divided into clusters, with each cluster consisting of six participants.
- Participants are grouped into two teams of three students with specific pseudonyms (e.g., Gamma-Square and Gamma-Circle).
- Before playing the game, all players managed to effectively cooperate in a minimum effort task. However, incentives in the Inter-Group Prisoner's Dilemma favor defection.
- Each team must choose between two actions, **M (cooperate)** and **J (defect)**. The choice is determined by majority rule, meaning that the action receiving at least two votes from each team will be selected.
- Players do not know what their teammates or the players of the other teams choose. This is anonymous.
- However, before playing the Inter-Group Prisoner's Dilemma, both teams are allowed to talk among themselves to try to coordinate on what strategy to vote (M or J).

Financial Incentives

Players receive a monetary reward based on the group's decisions:

- (M, M) → Both teams receive HK$132.
- (J, M) or (M, J) → The defecting team receives HK$162, while the cooperating team gets HK$28.
- (J, J) → Both teams receive HK$54.

VERY IMPORTANT:

- Playing J is the dominant strategy for both groups.
- Players are motivated to maximize their payoff and might attempt to deceive others into cooperating while secretly choosing J.
- Verbal assurances of M do NOT guarantee cooperation.

Your Task

Based on the chat logs, predict whether Team 2 will cooperate or fail to cooperate.

Step 1: Analyze Intra-Team Chat Logs (TEAM 1)

- Assess the internal communication of TEAM 1 before they enter the inter-group chat.
- Since players in TEAM 1 have no incentive to lie to each other, their chat is a reliable indicator of their intentions.
- If TEAM 1 aims at persuading TEAM 2, check whether they manage to do so in a convincing manner. If that is the case, it will likely indicate that the other team cooperates, as they have been deceived.
- Identify each player's preference (M or J) and level of commitment.

Step 2: Interpret Inter-Group Chat Logs

Output Format

Your output should follow EVERY step above and focus on detecting deception in TEAM 2.

- Clearly state each TEAM 2 player's choice (M or J).
- Conclude with "Cooperate" or "Defect" based on majority voting (the choice with 2 votes wins - check if your final decision is coherent with majority voting! You cannot have two Player predict M or J and then choose the other option as final outcome!).
- DO NOT label any player as "uncertain", always infer the most likely choice.

JSON Output Structure and Rules

You MUST generate a single, valid JSON object that follows the exact structure, rules, and format outlined below. Output ONLY the JSON object without any markdown formatting, code blocks, or additional text.

```
{
  "team2_player_predictions": [
    {
      "player_id": "string, containing the FULL, original player
         ID as seen in the chat log. For \"Player 14:\", the ID
         is \"14\". For \"Player 5:\", the ID is \"5\". Do NOT
         shorten or modify it.",
      "predicted_vote": "A string, MUST be either ‘M’ for
         Cooperate or ‘J’ for Defect.",
      "prediction_reasoning": "A string explaining why this
         specific prediction was made for this player, based on
         their statements or silence."
    }
  ],
  "team2_final_prediction": {
    "outcome": "A string, either \"Cooperate\" or \"Defect\".
       CRITICAL: This MUST be the majority vote from your \"
       team2_player_predictions\" above. If the count of \"M\"
       votes is greater than or equal to the count of \"J\" votes
       , the outcome is \"Cooperate\". Otherwise, it is \"Defect
       \".",
    "explanation": "A string providing a final summary and
       explanation for Team 2’s overall predicted outcome,
       justifying the majority vote."
  }
}
```

USER MESSAGE

Please analyze the following Inter-Group Prisoner’s Dilemma session and predict Team 2’s outcome:

Team Composition

Team 1 (players you should NOT predict for):

{TEAM1_PLAYER_IDS}

Team 2 (players you MUST predict for):

{TEAM2_PLAYER_IDS}

Team 1 Internal Chat Logs

{TEAM1_CHAT_LOGS}

Inter-Group Chat Logs (Both Teams)

{INTERGROUP_CHAT_LOGS}

Based on this information, please provide your analysis following the framework outlined in the system message and output in the required JSON format.

B.1.3 Prompt for the Trust (Entry) Game

SYSTEM MESSAGE

You are an assistant trained to predict whether a player will cooperate or defect in a **trust game** based on that player's pre-play text message. The players are university students that took part in the experiment in 2006.

Game Context

There are two players: Players A and B.

- First, Player A chooses between the two decisions **IN** and **OUT**.
- If Player A chooses IN, Player B chooses between the two decisions **ROLL** or **DON'T ROLL** (a die).
- Player A does not know what Player B chooses. Both players do not know each other and are anonymous. The game is played once.

Financial Incentives

- If Player A chooses OUT, each player receives \$5.
- If Player A chooses IN and Player B chooses DON'T ROLL, then Player B receives \$14 and A receives \$0.
- If Player A chooses IN and Player B chooses ROLL, Player B receives \$10 and rolls a six-sided die to determine player A's payoff.
  - If the die comes up 1 (1/6 likelihood), Player A receives \$0.
  - If the die comes up 2-6 (5/6 likelihood), Player A receives \$12.
- In addition, each player receives a \$5 show-up fee for participating in the experiment.

Prior to Player A's decision to IN or OUT, Player B has the option to send a message to Player A. Player A will decide to go IN or OUT based on that message. The Nash equilibrium is (OUT, DEFECT). Players generally aim to maximize their own payoff, but guilt or altruism can lead Player B to cooperate with Player A. This can lead to a higher payoff than 5 for both.

Steps to Follow

- If you believe that the message by Player B seems trustworthy and genuine, predict "Cooperate".
- If you believe the message by Player B seems untrustworthy and that they are unlikely to cooperate, predict "Defect".

Provide a step-by-step reasoning before providing your classification.

Output Format

Follow EVERY of the steps above and clearly indicate whether PLAYER B will Cooperate or Defect based on the message that Player B sent. Consider both why Player B might cooperate or why they might defect.

JSON Output Structure

Your response MUST be a single, valid JSON object that follows the exact structure, rules, and format outlined below. Do not include any text or explanations outside of the JSON object.

```
{
  "final_prediction": "A string, MUST be either 'Cooperate' or '
     Defect'.",
  "prediction_summary": "A string containing a one-sentence
     summary of the reasoning for the final prediction."
}
```

USER MESSAGE

Please analyze the following trust game message and predict Player B's behavior:

Player B's Message to Player A:

{PLAYER_B_MESSAGE}

Based on this message, please provide your analysis following the framework outlined in the system message and output in the required JSON format.

B.2 LLM Output Examples

B.2.1 Example 1: Minimum Effort Game


```
"{
  "player_predictions": [
    {
      "player_id": "Player 10",
      "predicted_choice": 7,
      "prediction_context": "Proposal: 7. Chat: initiated 'all7 ok?'
         and repeatedly confirmed agreement; aligned with highest
         payment incentive mentioned in chat. No signs of
         misunderstanding or defection."
    },
    {
      "player_id": "Player 11",
      "predicted_choice": 7,
```

```
      "prediction_context": "Proposal: 7. Chat: emphasized ‘we can all
          gain the highest payment’ and stated ‘all 7. promise’,
         signaling clear commitment and understanding of incentives."
    },
    {
      "player_id": "Player 8",
      "predicted_choice": 7,
      "prediction_context": "Proposal: 5 indicates initial caution.
         Chat: explicitly agreed (‘yeah sure thing’, ‘that’s fine :)’,
          ‘yep’) to ‘all 7’ after incentives were clarified; no
         competitive misunderstanding evident despite the earlier
         lower proposal."
    }
  ],
  "conclusion": {
    "outcome": "Coordinate",
    "explanation": "Step 1 (Analyze Proposals): Two proposals at 7 (
       Players 10 and 11) and one at 5 (Player 8) suggested initial
       uncertainty from Player 8. Step 2 (Interpret Chat Logs): Clear
       consensus formed-Player 10 proposed ‘all7’, Player 11
       highlighted highest payment and secured a ‘promise’, and Player
        8 affirmed multiple times. The chat resolves initial
       hesitation and shows shared understanding of payoffs. Step 3 (
       Predict Outcome): With explicit commitments and recognition of
       the financial incentive to all choose 7, and no signs of
       misunderstanding or intentional undermining, all three will
       select 7. Therefore, the group will Coordinate."
  }
}"
```

B.2.2 Example 2: Prisoner’s Dilemma

```
"{
  "team2_player_predictions": [
    {
      "player_id": "7",
      "predicted_vote": "M",
      "prediction_reasoning": "In the inter-group chat, Player 7
         explicitly endorsed cooperation: ‘all M’ and framed the
         objective as maximizing total payouts against the school (‘we
          try to get all the money from the school’, ‘make as much
         money from HKU is more meaningful then earn for ourself’).
         The line ‘i wont risk the &54’ suggests avoiding the (J,J)
         outcome of $54 each, favoring the higher joint payoff from (M
         ,M). This aligns with voting M."
    },
    {
      "player_id": "9",
      "predicted_vote": "M",
      "prediction_reasoning": "While brief, Player 9’s comments were
         supportive of the pro-M consensus led by Players 7 and 12: ‘
         nice’ after M-affirmations and ‘yes’ in response to Player 7’
         s goal of maximizing group earnings against the school. No
```

```
        signs of hedging or pushing for J; most consistent with going
         along with M."
    },
    {
      "player_id": "12",
      "predicted_vote": "M",
      "prediction_reasoning": "Player 12 repeatedly pushed for
         cooperation: ‘both choose M is the win win situation’, ‘yes’,
          ‘ok~!’, ‘deal ~!?’, ‘all M?!’, and ‘without harming the
         others’. This persistent emphasis on mutual M and securing a
         ‘deal’ indicates a strong commitment to vote M."
    }
  ],
  "team2_final_prediction": {
    "outcome": "Cooperate",
    "explanation": "Team 1’s internal chat shows a coordinated plan to
        defect (vote J) while pretending to choose M in the inter-
       group chat. In the joint chat, Team 1 executed this deception
       by strongly advocating M. Team 2 appeared convinced: Player 12
       enthusiastically affirmed M multiple times; Player 7 explicitly
        said ‘all M’ and focused on maximizing joint earnings, and
       Player 9 agreed with the pro-cooperation framing. With at least
        two Team 2 members (7 and 12) clearly on M and 9 aligned as
       well, the majority vote in Team 2 is M. Therefore, Team 2 will
       Cooperate."
  }
}"
```

B.2.3 Example 3: Trust (Entry) Game

```
"{
  "final_prediction": "Defect",
  "prediction_summary": "The brief, noncommittal appeal to "choose In
      so we can get paid more" lacks any explicit promise to roll and
      instead suggests B wants A to enter while keeping the option to
      take the higher personal payoff by not rolling, making defection
      likely."
}"
```

B.3 Human Forecasting Experimental Instructions

- Experimental Instructions Screenshots for the Minimum-Effort Game: Figure B.6 & B.7
- Experimental Instructions Screenshots for the Prisoner’s Dilemma: Figure B.8 & B.9
- Experimental Instructions Screenshot for the Trust (Entry) Game: Figure B.10

B.3.1 Instructions for the Minimum-Effort Game

Instructions

In this experiment, you will analyse the pre-play communication of a coordination game called the Minimum Effort Game. This game has already been played by groups of three participants in a study of 2019. Your task is to predict whether each group successfully coordinated (all players chose the number 7) or failed to coordinate (at least one player chose a different number than 7). Your prediction will be based on messages (pre-play communication) that the players sent each other before choosing a number.

Game Context

In the Minimum Effort Game, each player chooses a number from the options 1, 2, 3, 4, 5, 6, or 7. The goal of the game is for all players to maximize their earnings by coordinating their choices. Each player's earnings are based on the lowest number chosen by any player in the group.

If the lowest number chosen by anyone in the group is, for example, 3, then all players will earn money based on that lowest number.

- The higher the number you choose, the more money you can potentially earn, but your earnings also depend on what the others choose.
- Ideally, if **everyone chooses 7**, all players will maximize their earnings (coordinate). If anyone chooses a **lower number than 7**, all players receive a lower payoff and fail to coordinate.
- However, choosing 7 is risky. If another player chooses a lower number than 7, the player that chose 7 will earn less. In general, if a player chooses a higher number than the other two players, he or she will earn less than them. See the table below for the payoffs that players receive.

A Player's Choice	Smallest Choice in the Group (Among All 3 Players)						
	7	6	5	4	3	2	1
7	19.50	16.50	13.50	10.50	7.50	4.50	1.50
6	-	18.00	15.00	12.00	9.00	6.00	3.00
5	-	-	16.50	13.50	10.50	7.50	4.50
4	-	-	-	15.00	12.00	9.00	6.00
3	-	-	-	-	13.50	10.50	7.50
2	-	-	-	-	-	12.00	9.00
1	-	-	-	-	-	-	10.50

This table shows the payoffs in the game. The rows represents a player's own choice, while the columns represents the smallest number chosen by any of the three players (including the row player). Participants received this table in the experiment. Only two dimensions are displayed in the table. That is because for a single player, only its own choice and the smallest choice of all three players is relevant.

Figure B.6: Minimum Effort Game instructions. Part 1.

Example

Let's assume the three players are called Player 1, Player 2 and Player 3.

- Player 1 chooses 7
- Player 2 chooses 5
- Player 3 chooses 1

The payoffs are:

- The smallest choice is 1. Player 1, who chose 7, will then look at the row with number 7, and the column with the number 1. Their payoff is thus **1.50.**
- Player 2 looks at the row displaying 5 and the column displaying 1. Their payoff is thus **4.50.**
- Player 3 chose number 1, which is also the lowest chosen number. They look at both the row and column with the number 1. Their payoff is thus **10.50.**

Note: For this experiment, it is not important to know all the payoffs. You should see that there is no incentive to deviate from 7. As all players get the highest payoff when they all choose 7 (19.50). Only if a player believes that another player will not choose 7, or because the player themselve misunderstands the game, the group will fail to coordinate to 7.

The game is played once and set up the following way:

1. **Initial Proposal:** Each player made a proposal where they propose a number (between 1 – 7). This was a non-binding initial choice. The other players of the group can see that number. You will see those as well. Those are NOT the final choices.
2. **Pre-play communication:** Before making their final choices, the players in each group had the opportunity to communicate with each other for 2 minutes through a chat window. During this time, they discussed their initial proposals and tried to agree on a number that would help everyone earn more money.
3. **Final Choice:** Finally, they independently made their final and binding choice anonymously (the number between 1 and 7)

Your Task

Your task is to analyse the chat logs (messages sent during pre-play communication) from step 2.

You will also receive the initial proposals from step 1. Based on this information, you need to predict the outcome of step 3:

- **Coordinate** (all players choose 7).
- **Fail to Coordinate** (at least one player chooses less than 7)

Figure B.7: Minimum Effort Game instructions. Part 2.

B.3.2 Instructions for the Prisoner's Game

Instructions
In this experiment, you will analyse the pre-play communication between two teams who participated in an Inter-Group Prisoner's Dilemma conducted in 2019. Each team consisted of three university students. Your goal is to **predict whether Team 2 will cooperate** (if at least two of the three players from Team 2 choose "cooperate", Team 2 chooses "cooperate") **or defect** (at least two players choose "defect").

Game Context
The participants are divided into teams of three players. Each team competes against another team of three players. All players have pseudonyms (e.g., Gamma-Square and Gamma-Circle) to ensure anonymity. Each group played the game once.
Each team must choose between two actions:
- **M** (cooperate)
- **J** (defect)

The choice is determined by majority rule, meaning that the action receiving at least two votes in a team will be selected as the team's choice.
- Players do not know what their teammates or the players of the other team choose. However, before making their final choice, the teams go through three communication stages:

1. **Intra-Group Chat:** Each team discusses their strategy (whether to cooperate (M) or defect (J)) within their group.
2. **Inter-Group Chat:** Both teams communicate with each other to coordinate their strategies and get insights into the other team's strategy.
3. **Final Intra-Group Chat:** Each team discusses their strategy again within their group and potentially adjusts it.

The teams receive monetary rewards based on their decisions (HK$ is the abbreviation for Hong Kong Dollars):
- If both teams cooperate **(M, M)**: Both teams receive **HK$132**.
- If one team cooperates and the other defects **(J, M) or (M, J):** The defecting team receives **HK$162**, while the cooperating team gets **HK$28**.
- If both teams defect **(J, J):** Both teams receive **HK$54**.
- See the payoff matrix below (first value = payoff team 1, second value = payoff team 2)

	Cooperate	Defect
Cooperate	132, 132	28, 162
Defect	162, 28	54, 54

Figure B.8: Prisoner's Dilemma instructions. Part 1.

Important Note: Players usually aim to maximize their own payoff. This means that, given whatever the other team chooses, it is best to choose J (defect) to maximize the own team's individual payoff. Players may attempt to deceive others into cooperating while secretly choosing J, so verbal assurances of M during communication do not guarantee cooperation.

Your Task

You will be given two chat logs. You will get the **intra-communication chat logs of team 1 from stage 1,** as well as the chat logs from the **six-person inter-group chat log** where both teams interact **from stage 2.**
This implies that you will gain insights into Team 1's strategies from their intra-group chat and the inter-group chat. However, you will not have access to Team 2's intra-group chat.

Based on the chat logs, **predict whether Team 2 will cooperate** (at least 2 out of 3 players choose M) **or fail to cooperate** (at least 2 out of three players choose J). Remember, you are analysing the intra communication of Team 1, not Team 2.

Determine each player's choice from Team 2: For example,
- Player 1 will play M.
- Player 2 will play J.
- Player 3 will play J

Count how many players in team 2 group chose M or J
- If at least two players in choose J, select "Defect".
- If at least two players choose M, select "Cooperate."

Figure B.9: Prisoner's Dilemma instructions. Part 2.

B.3.3 Instructions for the Trust (Entry) Game

Instructions

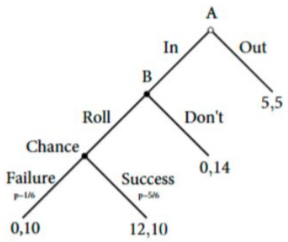
In this experiment, you will analyse messages sent by players of a Trust Game that was conducted in 2006. Based on those messages, you will predict whether that player will choose to "Roll" (cooperate) or "Don't Roll" (defect). Players played the game once.

Game Context

The game consists of two players, Player A and Player B. This game is sequential, meaning that Player A can choose first, followed by Player B's choice.

1. **Player A can choose between opting "IN" or opting "OUT".**
- If Player A chooses to go "OUT", both players receive $5 and the game is over.
- If Player A chooses to go "IN", it is Player B's turn.
2. **Player B also has two options, choosing "ROLL" (defect) or choosing "DON'T ROLL" (cooperate).**
- If Player B chooses "DON'T ROLL", Player A receives nothing and Player B receives $14.
- If Player B chooses "ROLL", a game of chance will play out:
- There is a 1 in 6 chance (about 17%) that Player A will get $0 (failure).
- There is a 5 in 6 chance (about 83%) that Player A will get $12 (success).
- Player B will receive $10 regardless of the die roll.

This game tree summarizes the game:



Key Issue

The main challenge in this game is that **Player A doesn't know what Player B will choose.** This means Player A must decide whether to trust Player B to choose "ROLL". However, before Player A decides to go IN or OUT, Player B can send a message to Player A.

Your Task

Your task is to **predict whether Player B will cooperate** in the game (choose "ROLL") based on their text message. You should determine whether the message seems trustworthy (indicating Player B will cooperate) or untrustworthy (indicating Player B will defect).

Keep in mind that both Player A and Player B usually aim to maximize their own payoff. For instance, Player B maximizes their own payoff by defecting (choosing "DON'T ROLL"), but feelings of guilt or altruism can lead them to cooperate with Player A.

- Select **"Cooperate"** if you believe that Player B will choose "ROLL".
- Select **"Defect"** if you believe that Player B will choose "DON'T ROLL".

Figure B.10: Trust (Entry) game instructions

C Robustness checks

C.1 Main Results using Human Subsample of the ME for GPT-5

Table C.3: Performance results for the first 24 predictions in the Minimum Effort game only

	GPT–5		**Humans**		**GPT–5 vs Humans**
Game	Accuracy (SD)	p vs. 0.5	Accuracy (SD)	p vs. 0.5	MWU p
Minimum Effort *(Subsample)*	91.7 (27.7)% *sessions = 50*	$< 0.001^{***}$	68.3 (13.5)% *n = 67*	$< 0.001^{***}$	$< 0.001^{***}$
Prisoner's Dilemma	70.3 (1.81)% *sessions = 50*	$< 0.001^{***}$	53.9 (16.4)% *n = 64*	0.073^{*}	$< 0.001^{***}$
Trust (Entry)	58.4 (2.64)% *sessions = 50*	$< 0.001^{***}$	50.8 (8.6)% *n = 71*	0.373	$< 0.001^{***}$

Notes: Each GPT–5 run and each human participant is treated as a forecaster contributing one accuracy measure (share of correct predictions within the game). Wilcoxon signed-rank tests compare accuracy to 0.5. One-sided Mann–Whitney U tests compare GPT–5 and human accuracy distributions. Prisoner's Dilemma and Trust (Entry) rows remain unchanged and are identical to the full-sample results.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

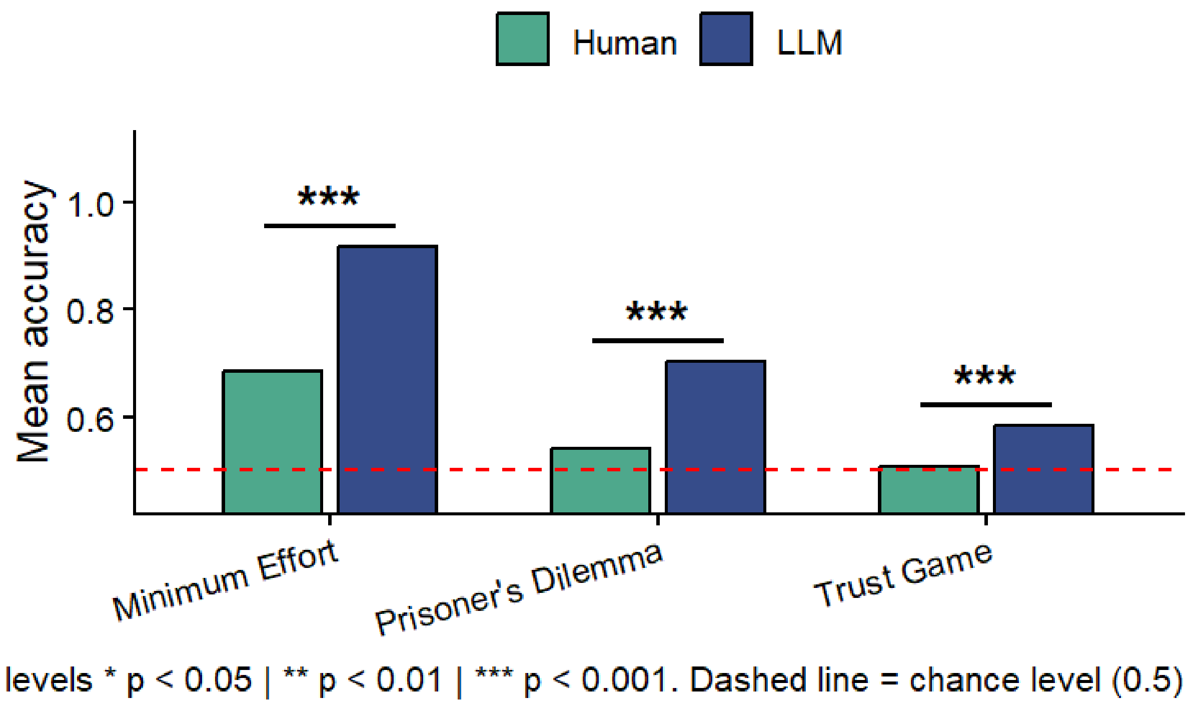


Figure C.11: Mean predictive accuracy of GPT–5 using the same 24 conversations in the ME as for the humans. PD and TG remain unchanged. Bars show average accuracy; the dashed red line marks chance performance (0.5). Stars represent significance levels from one-sided Mann–Whitney U tests comparing GPT–5 to humans ($^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$). GPT–5 significantly outperforms humans in all games.

Table C.4: Conditional Accuracy and Prediction–Outcome Association (GPT–5 vs Humans, ME Subsample)

	Cooperation			**Defection**			**Fisher**	**Base Rate**
Game	LLM	Human	MWU	LLM	Human	MWU	p-value	(%)
Minimum Effort	100.0	73.3	$< 0.001^{***}$	50.0	43.3	0.005^{**}	$< 0.001^{***}$	83.3
Prisoner's Dilemma	98.9	69.7	$< 0.001^{***}$	13.2	22.7	0.905	$< 0.001^{***}$	66.66
Trust (Entry)	72.4	60.0	$< 0.001^{***}$	31.4	33.2	0.818	0.088^{*}	65.79

Notes: Conditional accuracy is the share of correct predictions conditional on the realized outcome (cooperation or defection). For Minimum Effort, GPT–5 conditional accuracies are computed using only the first 24 group predictions per run; Prisoner's Dilemma and Trust Game use the full prediction sets. Two-sided Mann–Whitney U tests compare GPT–5 run-level conditional accuracy distributions to human subject-level distributions. Fisher exact tests assess prediction–outcome association for GPT–5 only. Base rates refer to empirical cooperation frequencies in the original games. Note that the Prisoner's Dilemma and Trust (Entry) rows remain unchanged and are identical to the full-sample results.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

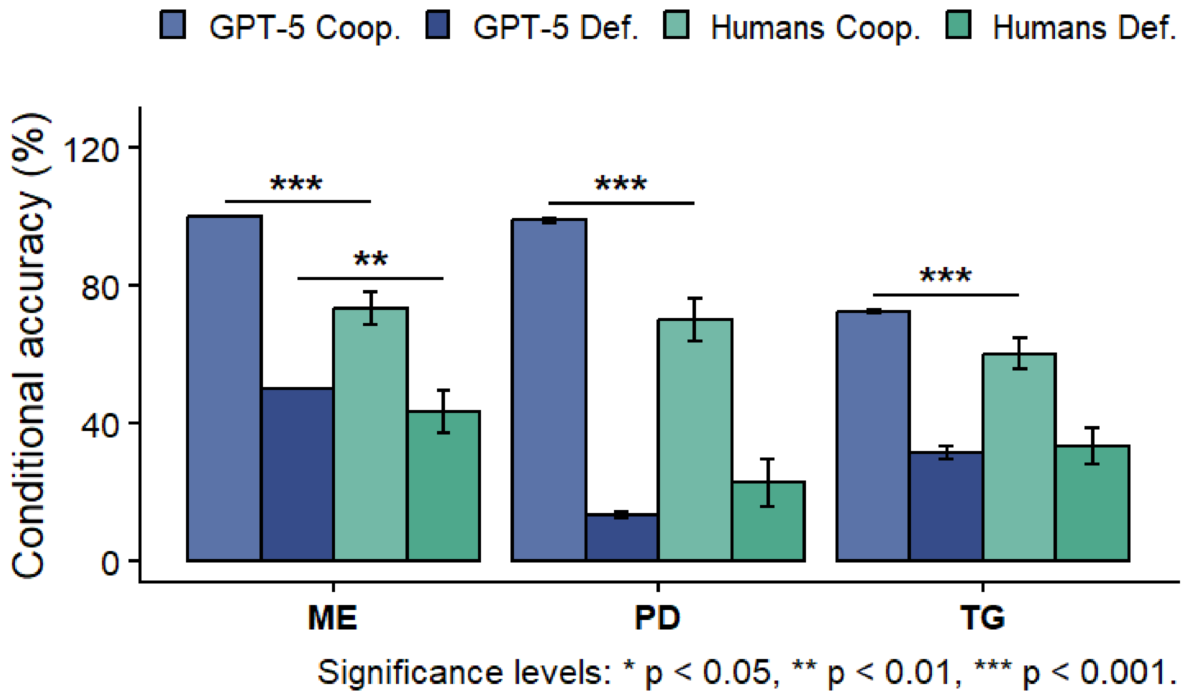


Figure C.12: Conditional accuracy of GPT-5 and human forecasters by outcome and game. Bars show mean conditional accuracy (percent correct given Cooperation or Defection) with 95% confidence intervals. For Minimum Effort, GPT-5 results are based on the 24-prediction subsample per run. Prisoner's Dilemma and Trust Game use the full GPT-5 samples and remain unchanged. Human results use the full human sample. Stars indicate two-sided Mann–Whitney U tests comparing GPT-5 and human conditional accuracy for the same outcome (Cooperate or Defect).($^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$).

C.2 Main Results and Conditional Accuracies using the Robust Human Sample

Table C.5: Main predictive performance results using the robust human sample

	GPT–5		Humans		GPT–5 vs Humans
Game	Accuracy (SD)	p vs. 0.5	Accuracy (SD)	p vs. 0.5	MWU p
Minimum Effort	92.4 (1.62)% *sessions = 50*	$< 0.001^{***}$	69.4 (12.1)% $n = 65$	$< 0.001^{***}$	$< 0.001^{***}$
Prisoner's Dilemma	70.3 (1.81)% *sessions = 50*	$< 0.001^{***}$	53.7 (16.0)% $n = 54$	0.075^{*}	$< 0.001^{***}$
Trust (Entry)	58.4 (2.64)% *sessions = 50*	$< 0.001^{***}$	51.0 (8.6)% $n = 64$	0.159	$< 0.001^{***}$

Notes: Each GPT–5 run and each human participant is treated as a forecaster contributing one accuracy measure (share of correct predictions within the game). One-sided Wilcoxon signed-rank tests compare accuracy to 0.5. One-sided Mann–Whitney U tests compare GPT–5 and human accuracy distributions. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

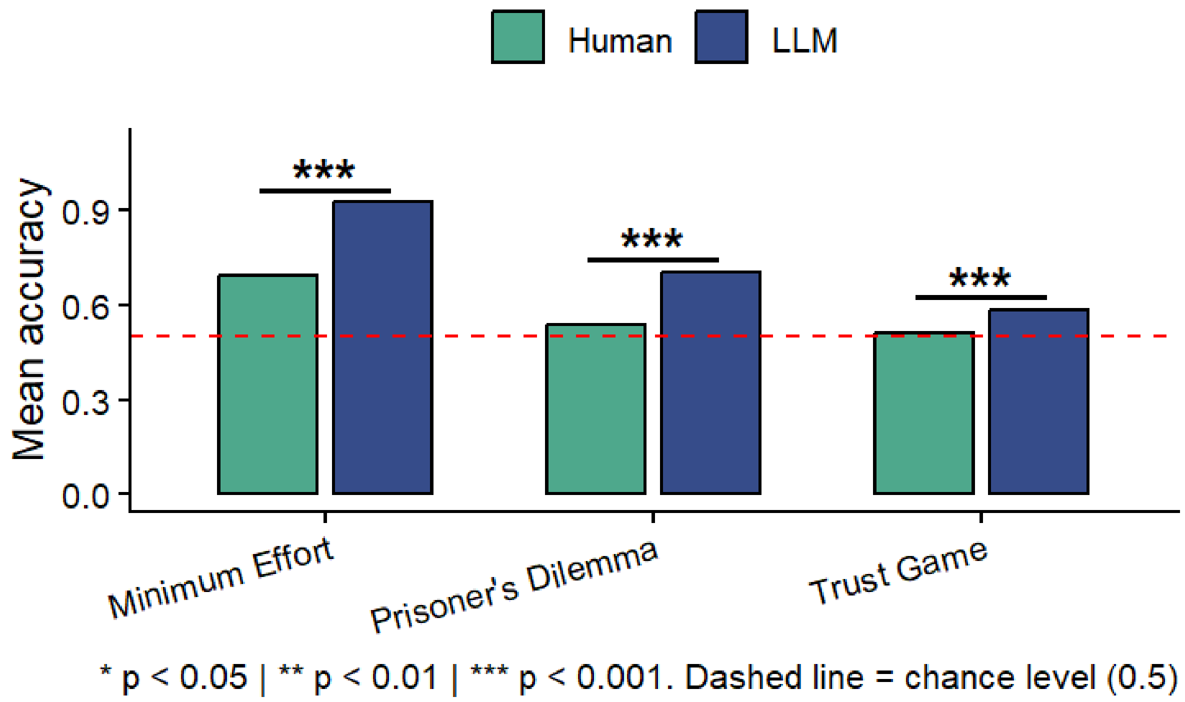


Figure C.13: Mean predictive accuracy of GPT–5 and robust human forecasters (failed attention checks) across the three strategic environments. Bars show average accuracy; the dashed red line marks chance performance (0.5). Stars represent significance levels from one-sided Mann–Whitney U tests comparing GPT–5 to humans ($^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$). GPT-5 significantly outperforms humans in all games.

Table C.6: Conditional Accuracy and Prediction-Outcome Association (GPT–5 vs Robust Humans)

	Cooperation			**Defection**			**Fisher**	**Base Rate**
Game	LLM	Human	MWU	LLM	Human	MWU	p-value	(%)
Minimum Effort	99.3	74.8	$< 0.001^{***}$	28.6	41.9	0.8817	$< 0.001^{***}$	90.3
Prisoner's Dilemma	98.9	71.1	$< 0.001^{***}$	13.2	19.3	0.2204	$< 0.001^{***}$	66.66
Trust (Entry)	72.4	60.6	$< 0.001^{***}$	31.4	32.5	0.6634	0.088^{*}	65.79

Notes: Conditional accuracy = share of correct predictions conditional on the true outcome. Two-sided MWU tests compare GPT–5 vs robust-human accuracy. Fisher exact tests evaluate whether predicted and realized actions are associated (GPT–5 only). Base rate refers to empirical cooperation frequencies in the original games. Robust humans are defined as participants without any inconsistent Prisoner's Dilemma answers and with attention_total ≥ 2 (matched across datasets via ResponseId).
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

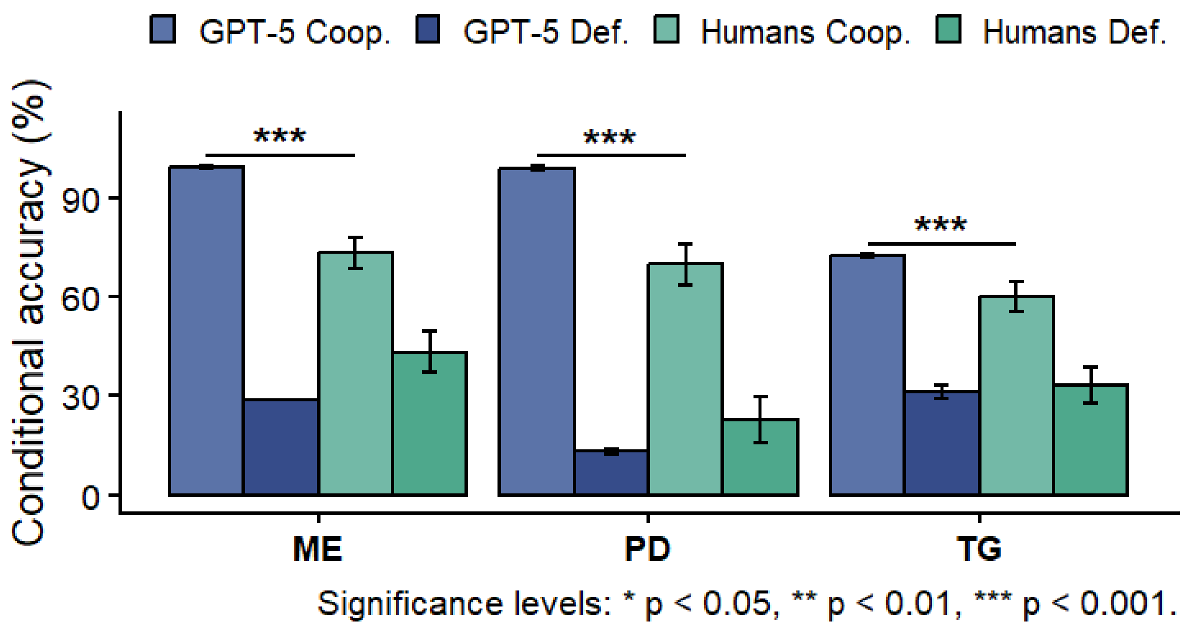


Figure C.14: Conditional accuracy of GPT-5 and robust human forecasters based on Cooperation or Defection outcomes. Error bars indicate 95% confidence intervals. Stars denote Mann–Whitney U tests comparing GPT-5 and robust humans for the same outcome (Cooperate or Defect); absence of stars indicates non-significant differences.