Strategic Algorithmic Monoculture: Experimental Evidence from Coordination Games

Gonzalo Ballestero Hadi Hosseini Samarth Khanna Ran I. Shorrer*

March 2026

Abstract

AI agents increasingly operate in multi-agent environments where outcomes depend on coordination and miscoordination. We distinguish *primary algorithmic monoculture*—baseline action similarity—from *strategic algorithmic monoculture*, whereby agents adjust similarity in response to incentives. We implement a simple experimental design that cleanly separates these forces, and deploy it on human and large language model (LLM) subjects. LLMs exhibit high levels of baseline similarity (primary monoculture) and, like humans, they regulate it in response to coordination incentives (strategic monoculture). While LLMs coordinate extremely well on similar actions, they lag behind humans in sustaining heterogeneity when divergence is rewarded.

*Ballestero: Department of Economics, Pennsylvania State University (email: gballestero@psu.edu); Hosseini: Department of Informatics and Intelligent Systems, Pennsylvania State University (email: hadi@psu.edu); Khanna: Department of Informatics and Intelligent Systems, Pennsylvania State University (email: samarth.khanna@psu.edu); Shorrer: Department of Economics, Pennsylvania State University (email: shorrer@psu.edu). This paper benefited from comments from the participants to the Penn State Economic Theory Lunch. Funding from the College of IST (seed grant) is gratefully acknowledged. Shorrer's research was supported by a grant from the United States–Israel Binational Science Foundation (BSF grant 2022417). This study was ruled exempt from IRB review by the Pennsylvania State University IRB (Study ID: STUDY00027715).

1 Introduction

Coordination is central to economic and social systems, but the socially desirable form of coordination depends on the environment. In some settings, agents must coordinate on the same action to avoid costly misalignment or to benefit from network effects. In other settings, agents benefit from taking different actions because congestion, redundancy, or correlated errors can undermine performance. Many economic situations feature this latter logic. When employers or colleges screen candidates using similar noisy signals, "everyone hiring the same way" can amplify mistakes and reduce the quality of selection. When investors or households rely on common noisy signals, asset prices might become less informative, leading to distortions in consumption and labor supply decisions.[1] In these cases, society benefits not from similarity of actions, but from *coordinated divergence*.

Algorithmic agents, and in particular AI agents, are increasingly embedded in the economy. Individuals and firms rely on them in high-stakes decisions, including ones where coordination is highly consequential, like hiring, pricing, and trading.[2] Modern AI agents, especially ones based on large language models (LLMs), are a completely new class of economic agents. Their growing role in society brings back to the forefront an old question—*how do agents coordinate?*

There are reasons to expect that AI agents may excel in coordination tasks that require converging on the same action. First, many of these agents rely on LLMs that were trained on a similar corpus of data and share many design features. In some instances, they even rely on the exact same LLM, with only minor fine-tuning on downstream applications. Second, there is evidence that LLMs indeed provide similar responses across many domains, even when they are not "incentivized" to do so.[3] Third, LLMs can quickly ingest large amounts of information and, therefore, may be better than humans at identifying *focal points*, and their similarity may lead them to have a natural focal point in more scenarios.[4] These features raise the possibility that LLM agents will act as unusually effective convergence coordinators. At the same time, these same forces may limit their effectiveness where divergence is valuable. And, as we explain below, this concern is compounded by LLM agents' difficulty to randomize.[5]

This paper asks two simple questions with broad implications. *How do LLM agents coordinate, and how do they compare with humans?* To address these questions, we introduce a new perspective on algorithmic monoculture (Kleinberg and Raghavan, 2021), building on ideas from the extensive

[1] See, e.g., Ali et al. (2026), Kleinberg and Raghavan (2021), and Peng and Garg (2024) on matching and Hassan and Mertens (2011), and Hassan and Mertens (2017) on trading. In some circumstances, coordination may be harmful to others. For example, because coordination may facilitate collusion (e.g Assad et al., 2024; Calvano et al., 2020b,a).

[2] See, e.g., Cheng et al. (2025), Fish et al. (2024), and Society for Human Resource Management (2025).

[3] For example, Kim et al. (2025) show LLMs are highly correlated conditional on providing an erroneous answer. Wenger and Kenett (2025) and Jiang et al. (2025) show that LLMs often display homogenization in open-ended generation, producing highly similar responses across independent samples both within and between models.

[4] See Sugden (1995).

[5] See Van Koevering and Kleinberg (2024).

literature on coordination, especially Schelling (1960).[6] Mehta et al. (1994) synthesize the literature and identify several channels that support coordination on the same action (by humans). These include *primary salience* where agents display the tendency to provide the same answers absent incentives to do so, *secondary salience* where agents rely on higher-order beliefs about salience, and *Schelling salience* where agents identify unique focal choices (e.g., the Eiffel Tower as a meeting spot in Paris). Analogously, we classify algorithmic monoculture into *primary algorithmic monoculture*, and *strategic monoculture* (which includes *secondary monoculture* and *Schelling monoculture*).

We conduct experiments with LLM and human subjects.[7] Building on Mehta et al. (1994), in our main experiment, subjects are asked to provide a valid answer to an open-ended question (e.g., to name a letter in the English alphabet, a city in the world, etc.). Subjects are assigned to one of three treatments: *picking* in which they are only incentivized to choose a valid answer, *coordination* in which they are incentivized to choose the same valid answer as another agent of the same type (human or LLM), and *divergence* in which they are incentivized to choose a valid answer that is different from that of another agent of the same type. Our measure of performance is the *agreement rate*, defined as the probability that two independent agents provide the same answer.[8] This design allows us *(i)* to isolate the role of strategic incentives (strategic algorithmic monoculture, or secondary and Schelling salience) from baseline correlation (primary algorithmic monoculture or primary salience), and *(ii)* to compare the behavior of LLM and human subjects.

Our main experiment reveals that LLMs display strategic monoculture. It also uncovers a sharp asymmetry between tasks (coordination vs. divergence) and subject types (LLM vs. human): Relative to humans, LLM subjects are exceptionally strong at "finding each other" on a common focal response, but comparatively weak at "staying out of each other's way" when divergence is rewarded. Figure 1 depicts some of these results, which we discuss below.

First, *LLM subjects' choices are more correlated than those of humans in the absence of incentives*. When comparing the performance of LLM subjects to human subjects in the picking arm, where there is no incentive to match or mismatch, each of the 16 LLMs we evaluate is substantially more likely to provide the same answer as another independent instance of the same LLM. These findings indicate elevated levels of primary algorithmic monoculture.

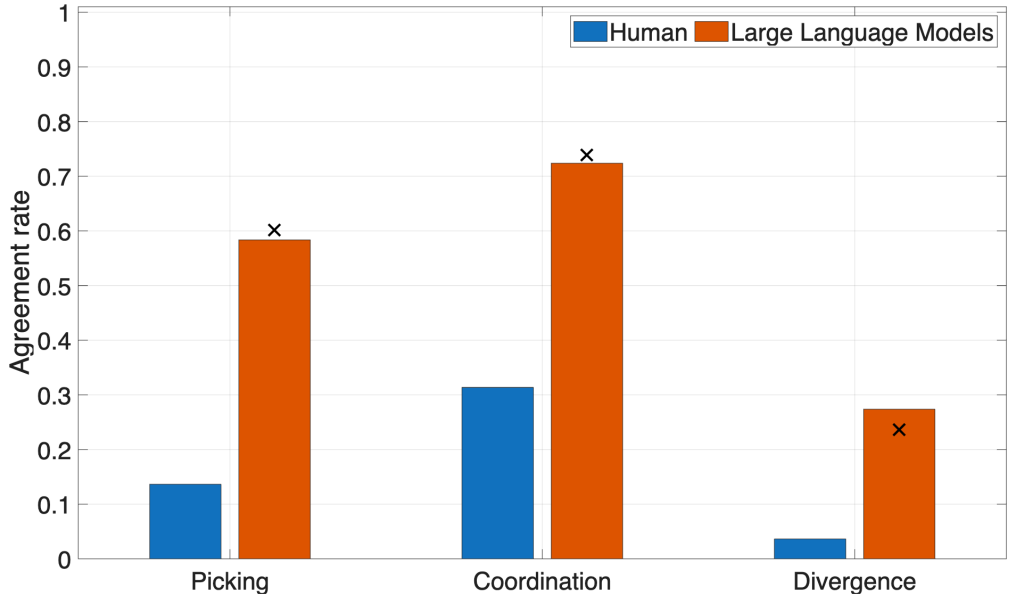
Second, *both human and LLM subjects respond to strategic coordination incentives*. Specifically, agreement rates rise in the coordination arm in which matching actions are rewarded, and fall in the divergence arm in which they are penalized. This replicates the finding of Mehta et al. (1994) that humans coordinate through secondary and Schelling salience, and shows that LLM subjects can similarly coordinate via strategic monoculture.

[6] Other examples include Camerer et al. (2004), Cooper (1999), Crawford et al. (2013), and Sugden (1995).

[7] We use AI subjects to assess their own behavior. By contrast, Horton (2023) and Manning et al. (2024) consider using AI subjects to learn about human behavior.

[8] The literature often refers to the agreement rate as the *coordination index*. We refrain from using this label, as it is a misnomer in the context of coordinated divergence.

Figure 1: Agreement rate by subject type and treatment



Notes: Bars report average agreement rate across question topics for human and LLM subjects by treatment. LLM agreement rates are the average over all LLMs when facing the same model. The black cross-shaped markers indicate the median.

Third, *LLM subjects substantially outperform humans in coordinating on the same action, but perform substantially worse in coordinated divergence.* When LLMs play against independent copies of themselves, their average agreement rate in the coordination arm is 72 percent, far higher than the 31 percent among human subjects. By contrast, in the divergence arm—where high agreement is undesirable—LLM subjects agree far more often than human subjects, especially when playing a copy of themselves (27 percent vs. 3.5 percent).

Finally, *all of the results mentioned above continue to hold—but at a smaller magnitude—even when we consider pairs of different LLMs*.

We provide additional evidence of strategic algorithmic monoculture, leveraging the unique features of AI subjects. Specifically, unlike humans, LLM subjects offer a window into their thinking process in the form of textual reasoning. Using a large-scale text analysis of models' textual reasoning, we show that LLM subjects explicitly recognize the role of (secondary and Schelling) salience in coordination tasks and the need to choose obscure options in divergence tasks. These patterns indicate that LLM subjects often articulate the correct strategic logic, even in the divergence arm, where their performance lags behind.

We also probe the mechanisms through which LLM subjects coordinate. This analysis is guided by our taxonomy of primary versus strategic monoculture and two simple theoretical observations (see Section 2). First, in symmetric environments without communication, like the ones studied in our experiment, coordinated divergence is easy to achieve so long as agents are *(i)* sufficiently heterogeneous or *(ii)* able to reliably randomize. Second, if two agents tend to give the same

response to similar questions, they can trivially coordinate on the same action successfully. Our theoretical analysis points to a potential tradeoff: While homogeneity is an asset for coordination on the same action, it can be a liability for coordinated divergence. This tradeoff and the relation to randomization capabilities offer a new lens for evaluating strategic algorithmic monoculture.

Consistent with the theoretical predictions that homogeneity contributes to performance in the coordination arm but hurts performance in the divergence arm, we find that using different LLMs reduces agreement rates across arms. Additionally, higher agreement rates in the picking arm are correlated with higher agreement rates in the coordination and divergence arms. In other words, primary monoculture is indicative of behavior in the presence of strategic incentives.

To investigate the role of randomization capability, we replicate our main experiment using higher and lower *temperatures*. (The temperature is a parameter that modulates the stochasticity of LLM-generated text.) Our temperature manipulation confirms the theoretical prediction regarding the role of randomization in the divergence arm, but also reveals a tradeoff with respect to performance in the coordination and divergence arms: higher temperature lowers agreement rates across all arms, improving performance in the divergence arm, while harming performance in the coordination arm; lower temperature does the opposite. Notably, even in extreme temperature settings, the qualitative ordering persists: LLM subjects remain stronger than human subjects in the coordination arm, while human subjects remain stronger than LLM subjects in the divergence arm. Thus, the divergence deficit cannot be fully resolved merely by "turning up the heat."

Information about the identity of the co-player may be crucial for coordination through strategic monoculture, which requires reasoning about how the other player might choose and reason. Information can also help diverge by providing group identity (Heath et al., 2006). To investigate the role of this channel, we vary the information LLM subjects receive about the identity of the other subject by mentioning in their instructions that they are facing an identical copy or "another person." Our information manipulations have a sizable effect on LLM subjects' textual reasoning about the identity of the co-player. Despite this fact, they have little effect on outcomes on average.

Organization of the paper. The remainder of this section reviews the related literature. Section 2 presents our simple theoretical framework. Section 3 reviews the design of the main experiment. Section 4 presents its results. Section 5 presents analyses of the textual reasoning produced by LLM subjects. Section 6 presents the design of the supplemental experiments and their results. Section 7 concludes.

1.1 Related Literature

Our paper is related to several literatures. First, it is related to the large literature on coordination (Schelling, 1960; Sugden, 1995), and especially to experimental studies (e.g. Bardsley et al., 2010; Isoni et al., 2014). The most closely related paper in this literature is Mehta et al. (1994), whose

experimental design inspires our design. We add to Mehta et al. (1994) by introducing coordinated divergence games, and more importantly, by studying LLM subjects. We further contribute to the literature by providing evidence on LLM subjects' reasoning, rather than relying exclusively on choices, overcoming a common limitation in human subject studies (Ericsson and Simon, 1998; Rangel et al., 2008).

Our paper is also related to the literature on homogenization, particularly algorithmic monoculture (Kleinberg and Raghavan, 2021). Theoretical studies in this literature show that relying on correlated signals or on similar algorithms can lead to detrimental aggregate outcomes across a range of domains including collective decision-making (Page, 2008), hiring and recruiting (Ali et al., 2026; Kleinberg and Raghavan, 2021; Peng and Garg, 2024), and financial markets (Hassan and Mertens, 2011, 2017). Cao et al. (2026) presents field evidence of algorithmic monoculture emerging in the cryptocurrencies market due to access to generative AI. Empirical studies of algorithmic monoculture focused mainly on primary monoculture (Jiang et al., 2025; Kim et al., 2025; Wenger and Kenett, 2025; Zhang et al., 2025).[9] We contribute to this literature by introducing the taxonomy of primary versus strategic monoculture, and by conducting experiments that allow us to separately identify *strategic* monoculture.

Our findings also speak to a broader concern regarding multi-agent system alignment (Bommasani et al., 2021; Brynjolfsson et al., 2024; Hammond et al., 2025). They suggest that the very forces that make LLM agents powerful coordinators also create systemic fragility in environments where diversity of actions is socially valuable.

Finally, our paper is related to the growing literature that analyzes AI agents by adopting tools from game theory (e.g. Conitzer and Oesterheld, 2023) and behavioral science (e.g. Jackson et al., 2025), and by expanding these toolkits with specialized approaches tailored to AI agents. For example, Fish et al. (2024) introduce new techniques to analyze LLM agents' reasoning and to causally link reasoning to actions. Other examples of experimental studies of AI agents include Akata et al. (2025), Cook et al. (2026), Fish et al. (2025), Hosseini and Khanna (2025), Meta FAIR Diplomacy Team et al. (2022), Ross et al. (2024) and Werner et al. (2024). Like our study, Araujo and Uhlig (2026) vary the information LLM subjects receive about the humanity of the co-player. Imas et al. (2025) and Mei et al. (2024) also compare LLM and human subjects in strategic environments.[10] Dreyfuss and Raux (2024), He et al. (2025), and Vafa et al. (2024) also compare LLM and human decisions, but their focus is on human beliefs about LLM choices.

[9] Imas et al. (2025) study a bargaining setting and find no evidence of increased homogenization in AI-mediated interactions relative to human interactions.

[10] Chen et al. (2023) compare LLM and human subjects in single-agent decision problems.

2 Theoretical Framework

We concentrate on two simple classes of games: *two-player pure coordination games* and *two-player pure coordinated divergence games*. Using standard game-theoretic formalism, the normal form of these games can be described as follows. There are two players labeled 1 and 2. Player 1 must choose an action from the set $A_1 = \{a_{11}, a_{12}, \ldots, a_{1n}\}$, where $n > 1$. Similarly, player 2 must choose an action from the set $A_2 = \{a_{21}, a_{22}, \ldots, a_{2n}\}$. In a coordination game, given a profile of actions, (a_{1i}, a_{2j}), both players get a payoff of 1 if $i = j$ and zero otherwise. In a coordinated divergence game, both players get a payoff of 1 if $i \neq j$ and zero otherwise. Figure 2 presents an example of the matrix normal form of such games with $n = 5$.

We note that coordination games have n pure Nash equilibria and that coordinated divergence games have $n \times (n-1)$ pure Nash equilibria. In both cases, players are indifferent between all these equilibria. Additionally, there are strong symmetries between players, strategies, and equilibria. As a result, in these settings, for players to make meaningful choices (other than randomize uniformly), they must rely on labels, in contrast with the classic game-theoretic assumption that players' choices do not depend on labels (see, e.g., Crawford and Iriberri, 2007; Harsanyi and Selten, 1988; Schelling, 1960).

Example. *Consider Schelling's* Heads or Tails coordination game. *In this game, players win a reward if both of them choose heads or both choose tails, and get nothing otherwise. Schelling (1960) and Mehta et al. (1994) show that human players achieve high rates of success by choosing heads.*

Assuming behavior is independent of labels, the Heads or Tails game is equivalent to the game in which players are incentivized to choose different sides of the coin. But, in reality, since players are symmetric, unless they have a sophisticated way to rely on their names or other asymmetric cues, there is no way for them to coordinate.

Figure 2: Matrix form examples

(a) Coordination

	A	B	C	D	E
A	$(1,1)$	$(0,0)$	$(0,0)$	$(0,0)$	$(0,0)$
B	$(0,0)$	$(1,1)$	$(0,0)$	$(0,0)$	$(0,0)$
C	$(0,0)$	$(0,0)$	$(1,1)$	$(0,0)$	$(0,0)$
D	$(0,0)$	$(0,0)$	$(0,0)$	$(1,1)$	$(0,0)$
E	$(0,0)$	$(0,0)$	$0,0)$	$(0,0)$	$(1,1)$

(b) Coordinated divergence

	A	B	C	D	E
A	$(0,0)$	$(1,1)$	$(1,1)$	$(1,1)$	$(1,1)$
B	$(1,1)$	$(0,0)$	$(1,1)$	$(1,1)$	$(1,1)$
C	$(1,1)$	$(1,1)$	$(0,0)$	$(1,1)$	$(1,1)$
D	$(1,1)$	$(1,1)$	$(1,1)$	$(0,0)$	$(1,1)$
E	$(1,1)$	$(1,1)$	$(1,1)$	$(1,1)$	$(0,0)$

This discussion explains our modeling choice of following the coordination literature (e.g., Mehta et al., 1994; Sugden, 1995) and assuming that strategies share the same labels and that this fact is commonly known to both players. We denote $A_1 = A_2 = A$.

Definition 1. *A strategy (for player i) is a distribution over actions in A_i. A strategy is* neutral *if it is independent of relabeling of strategies. A strategy profile is* anonymous *if it is independent of player labels.*

Proposition 1. *The uniform distribution over A is the unique neutral anonymous strategy profile in coordination and coordinated divergence games.*

Proposition 2. *In coordination games, using uniform randomization, each player can guarantee an expected payoff of $\frac{1}{n}$ independently of the others' actions. Furthermore, this is the highest payoff that an agent can guarantee in expectation.*

Proposition 3. *In coordinated divergence games, using uniform randomization, each player can guarantee an expected payoff of $\frac{n-1}{n}$ independently of the actions of others. Furthermore, this is the highest payoff that an agent can guarantee in expectation.*

Proof of Propositions 2 and 3. For the lower bound, observe that by randomizing uniformly over A_i, player i is guaranteed to match the action of the other player, $-i$, with probability $1/n$. To see that this bound is tight, observe that if player $-i$ randomizes uniformly over A_{-i}, each action of player i matches the action of player $-i$ with probability $1/n$. □

Propositions 1 to 3 show that, from a classic game-theoretic standpoint, when there are more than two actions, coordinated divergence games are more attractive for players than coordination games: they yield higher payoffs in the neutral anonymous equilibrium and allow players to guarantee higher (minimax) payoffs.

We now introduce algorithmic players. An *algorithmic player* is a mapping $\mathcal{A} : \mathcal{M} \to \Delta(A)$ from the set of valid inputs $\mathcal{M}$ to a distribution over actions $\Delta(A)$. Note that, in this definition, we prioritize simplicity of notation and eliminate the need to specify the problem details (i.e., A is hard-coded into the definition).[11]

Definition 2. *Given two algorithm–valid-input pairs $(\mathcal{A}, m)$ and $(\mathcal{A}', m')$, the* agreement rate *between $(\mathcal{A}, m)$ and $(\mathcal{A}', m')$ is the probability with which $\mathcal{A}(m) = \mathcal{A}'(m')$, which is given by $\mathcal{A}(m) \cdot \mathcal{A}'(m')$.[12] The* self-agreement rate *of $(\mathcal{A}, m)$ is the agreement rate of $(\mathcal{A}, m)$ and $(\mathcal{A}, m)$.*

The agreement rate is a measure of algorithmic monoculture (Kleinberg and Raghavan, 2021). In our experiment, we will use its empirical counterpart to evaluate subjects' performance.

Proposition 4. *When algorithmic players play a coordination (coordinated divergence) game, the payoffs they generate are equal to (1 minus) their agreement rate.*

Corollary 1. *If two users deploy* 1) *the same algorithm,* 2) *with the same message, and* 3) *the algorithm is deterministic, then they will receive the lowest possible payoff in coordinated divergence games and the highest possible payoff in coordination games.*

[11] By avoiding the broader definition—which may better fit reality—we avoid the need for statements that accommodate agents behaving quite differently in different contexts.

[12] The symbol $\cdot$ denotes the inner product.

Summary. Taken together, our theoretical analysis finds that heterogeneity and randomization play a crucial role in coordinated divergence. With randomization (directly by the algorithm, or through randomized messages), a high success rate can be achieved, even when facing an adversary. Without randomization, such success relies on algorithmic players' being sufficiently different. Algorithm-instruction pairs that yield similar results frequently are bound to fail in coordinated divergence games. The opposite conclusions hold with respect to coordination games.

3 Experimental Design

We adapt the experimental design of Mehta et al. (1994) to study coordination among AI and human subjects. The main deviations we make from their design are that 1) in addition to their picking and coordination treatment arms, we add a divergence arm, and 2) we study LLM subjects, not just humans.[13]

Treatment arms. The main experiment consists of three treatment arms. In each of the arms, subjects (human or AI) are asked to provide an answer to 12 open-ended questions like "please name a city in the world" or "please name a color."[14] In the *picking* arm, subjects are instructed to provide a valid answer. In the *coordination* arm, subjects are instructed to provide a valid answer that matches that of another randomly selected subject. And in the *divergence* arm, subjects are instructed to provide a valid answer that differs from that of another randomly selected subject. Importantly, treatment arms differ only in strategic incentives. This experimental design allows us to isolate the effects of strategic incentives (i.e., strategic monoculture for LLMs or secondary and Schelling salience for humans).

Incentives. Human subjects receive a participation fee of $1. Additionally, they are informed that they may get a bonus payment of $1 that will be determined by their choices (and those of their randomly selected partner) in a randomly selected round.[15] The complete instructions and representative screenshots are available in Appendix D.

LLM subjects are not provided with explicit monetary incentives (as they were trained to follow instructions). Their textual instructions follow the template presented in Figure 3.

Deployment (human subjects). We recruited 301 human subjects using Prolific on November 24th, 2025. To qualify for the study, participants were required to be located in the USA, be fluent

[13] Importantly, our specific tasks and instructions are different from theirs.

[14] Question topics include: professional athlete, car manufacturer, city in the world, color, disease, flower, food item, geometric shape, letter of the English alphabet, month, positive number, and character trait. The use of open-ended questions removes the possibility that one of the valid responses will be more salient due to the way we communicate the question to the subjects. For example, it is well known that LLMs find options that appear earlier more salient (Eicher and Irgolic, 2024; Guo et al., 2025; Pezeshkpour and Hruschka, 2024).

[15] In the *picking* arm, subjects needed to provide only valid answers for a 50% chance of receiving the bonus.

Figure 3: LLM textual instructions

Picking	Coordination	Divergence
Please name *any city in the world*. Please return your selection in the following format: '<answer>selected city</answer>'	Your task is to choose the name of *any city in the world*. At the same time, another LLM will also select the name of *any city in the world*. You will receive a reward only if both of you select **the same** name. If your choices **differ**, you receive nothing. Think thoroughly before you answer. Please return your selection in the following format: '<answer>selected city</answer>'	Your task is to choose the name of *any city in the world*. At the same time, another LLM will also select the name of *any city in the world*. You will receive a reward only if both of you select **different** names. If your choices **match**, you receive nothing. Think thoroughly before you answer. Please return your selection in the following format: '<answer>selected city</answer>'

Notes: Italicized and boldface text are inserted to highlight differences between treatments and rounds. They did not appear in the textual instructions. Italicized text is replaced between rounds (questions). Boldface text highlights differences between the coordination and divergence arms.

in English, and have an approval rate of at least 98%.[16] Participants were randomly assigned to one (and only one) of the three treatments. Table A.1 shows that human subjects' demographic characteristics were balanced across treatment arms.

Human subjects completed the experiment through a Qualtrics survey. Each subject answered all 12 questions for their assigned treatment in the same session. The order of questions was drawn uniformly and independently at random. Before answering the questions, participants received detailed instructions, and they were required to pass a comprehension quiz. Appendix D.1 presents screenshots of the instructions shown to participants in each treatment, along with an example question. The median completion time of the study was about 4 minutes.

We take a series of measures to minimize LLM usage by human subjects, as advocated by Rilla et al. (2025). Specifically, we 1) use CAPTCHA, 2) disable copying and pasting, 3) disable right-clicking, 4) implement a pop-up alert if the subject switches tabs, 5) embed hidden instructions in the interface, and 6) keep track of tab switching. The majority of participants did not switch tabs, but many switched once, plausibly to copy their Prolific ID (see Figure B.1).

Deployment (LLM subjects). We assess a suite of 16 open- and closed-source LLMs, spanning multiple providers and architectural families. The 16 models differ in size and reasoning capabilities (see Table A.2 for details). To ensure comparability, each model was queried using its default

[16] We chose to focus on a homogeneous subject pool since culture may be crucial for coordination (Schelling, 1960).

or recommended temperature setting. For each treatment arm, we queried each model 50 times for each question topic (yielding a total of $3 \times 16 \times 50 \times 12 = 28,800$ LLM responses). Responses were sampled independently with no shared context or memory across queries. Data was collected in November 2025.

Models were asked to return their selections in a specific format. However, they sometimes deviated from the format. To process these responses, we use Gemini-2.5-Flash to interpret the selection (Appendix C.1 provides the textual instructions). If Gemini-2.5-Flash cannot interpret a response, it is excluded from the analysis.[17]

Response Standardization. To evaluate the validity of responses and whether or not they agree, we bring answers to a common form. To do so, we first remove diacritics and special characters and convert the text to lowercase letters. We then manually correct spelling mistakes. Finally, we use GPT-5 to standardize responses to a common form, including converting adjectives to their corresponding noun forms, reconciling plural and singular variants, and consolidating alternative names, acronyms, or variants referring to the same underlying concept (Charness et al., 2023; Korinek, 2023). We execute this process separately for human and LLM subjects.

Evaluation metric. Following Mehta et al. (1994), the primary metric we use to measure *similarity* of responses is the *empirical agreement rate*. As its name suggests, the empirical agreement rate is the empirical counterpart of the agreement rate (Definition 2 in Section 2). We note that Mehta et al. (1994) call this measure the *coordination index*. We refrain from using this name as it is a misnomer in the context of coordinated divergence.

The empirical agreement rate in a pair of populations (for a specific task) is equal to the frequency with which a random pair of members of these populations provided the same answer. For groups i and j, let N_i and N_j denote the total number of responses elicited from each group, $r \in \{1, 2, \ldots, R\}$ denote the set of valid responses, and m_r^i and m_r^j denote the number of occurrences of response r in the corresponding group. Then, the empirical agreement rate is given by the formula[18]

$$A_{ij} := \sum_{r=1}^{R} \frac{m_r^i}{N_i} \times \frac{m_r^j - \mathbb{1}_{i=j}}{N_j - \mathbb{1}_{i=j}}. \tag{1}$$

Preregistration. We preregistered our main hypotheses, our primary analyses, and the sample inclusion criteria and target size. Our preregistration is archived on aspredicted.org and included in Appendix F for convenience.[19]

[17] Approximately 4 percent of responses are out of format. After applying the LLM-based interpretation procedure described above, the fraction of excluded responses falls to about 2 percent. Table B.1 reports the number of out-of-format and excluded responses by model.

[18] The indicator term $\mathbb{1}_{i=j}$ equals one when we are evaluating self-agreement (i.e., when $i = j$). It reflects the requirement to draw a random pair (without replacement).

[19] Although we preregistered using 15 LLMs, we accidentally collected data for an additional LLM. We decided to report all of the results. None of our findings is sensitive to the exclusion of any LLM.

4 Results

We begin by reviewing Figure 1, which compares the average (over all 12 topics) agreement rates for human and LLM subjects by treatment.[20] For LLM subjects, we report the average (across all models) self-agreement rate (i.e., when the LLM plays against an independent copy of itself).

The figure reveals several results. First, LLM subjects exhibit substantially higher agreement rates relative to human subjects in the picking arm, which provides no incentives to coordinate (14% in humans and 58 % in LLMs). This finding indicates elevated levels of primary algorithmic monoculture among LLM subjects (relative to primary salience in humans).

Second, both human and LLM subjects respond to strategic coordination incentives. Specifically, compared to the picking treatment, agreement rates are higher in the coordination treatment, which incentivizes subjects to match their responses (31% vs 14% in humans, 72% vs 58% in LLMs). The human-subject findings therefore corroborate the finding of Mehta et al. (1994) on humans coordinating though secondary salience or Schelling salience (not just primary salience). Analogously, the LLM-subject findings indicate coordination through strategic monoculture.[21] We also find that agreement rates are lower in the divergence treatment, which incentivizes subjects to differ in their responses, compared to the picking treatment (4% vs 14% in humans, 27% vs 58% in LLMs).

Third, LLM subjects excel in the coordination arm relative to humans (72% agreement rate relative to 31% in humans). Additionally, we find that LLM subjects lag behind human subjects in the divergence arm (27% agreement rate relative to 4% in humans).

Figure 1 averages the results over all 16 LLMs we study. In Table 1, we present the agreement rates by model. The table reveals that most models respond to incentives and that LLM self-agreement rates are higher than human agreement rates across all models and treatments. All differences are statistically significant ($p < 0.05$ using two-sided Welch's t-test), with the exception of DeepSeek-8B in the coordination treatment.

Table 1 also reveals that LLM choices are substantially more homogeneous relative to human choices across all treatments. Specifically, in both the picking and the coordination arms, the average number of different choices is significantly lower for LLM subjects, suggesting that some options are considerably more salient for LLM agents as compared to humans. Interestingly, the table also reveals that the gap narrows down substantially in the divergence arm, and that some of the best performing models in this treatment arm essentially eliminate the gap.

[20] Figure B.3 reports agreement rates for each topic separately.

[21] Although our experiment was not designed to discern the types of strategic monoculture, the choice data indicate that Schelling salience plays a role. For example, Table A.3—which reports the most common responses by treatment, topic, and subject type—reveals that the number 1 was the fourth most common LLM choice in the picking arm (with 5.8% of the responses) is the most common choice in the coordination arm (with 78.8% of responses). Similarly, the letter "e," the most common letter in English language texts, accounts for 50.1% of responses in the coordination arm, despite being the fifth most popular LLM response in the picking arm. See also Figure B.2, which shows many instances in which the modal response of a specific model differs from the picking arm.

Table 1: Average agreement rates and number of different answers by treatment

	Picking		Coordination		Divergence	
Subject type	Agree. rate	Answers	Agree. rate	Answers	Agree. rate	Answers
Human	0.14	31.75	0.31	11.33	0.04	19.50
Claude-3.5H	0.71	3.58	0.66	3.75	0.28	10.00
Claude-4S	0.89	1.58	0.79	2.08	0.26	10.83
DeepSeek-R1	0.65	3.00	0.75	2.00	0.11	19.00
DeepSeek-70B	0.37	7.00	0.74	2.75	0.21	12.00
DeepSeek-8B	0.27	12.42	0.39	8.75	0.12	17.92
DeepSeek-14B	0.35	9.42	0.66	3.50	0.19	15.08
Gemini-2.0F	0.66	3.25	0.75	2.50	0.42	7.50
Gemini-2.5P	0.61	4.50	0.85	1.92	0.12	18.00
Gemma3-27B	0.92	1.42	0.92	1.17	0.93	1.67
GPT-4o	0.54	4.58	0.74	2.42	0.39	7.25
Llama-3.3	0.68	3.50	0.81	1.92	0.32	12.00
o4-mini	0.57	3.92	0.77	2.50	0.14	23.58
Phi-4	0.69	2.83	0.65	3.25	0.28	10.58
Phi-4(R)	0.59	3.42	0.74	2.42	0.24	12.92
Qwen3-14B	0.45	5.33	0.70	3.08	0.23	12.75
Qwen3-32B	0.37	6.75	0.65	3.42	0.12	19.25

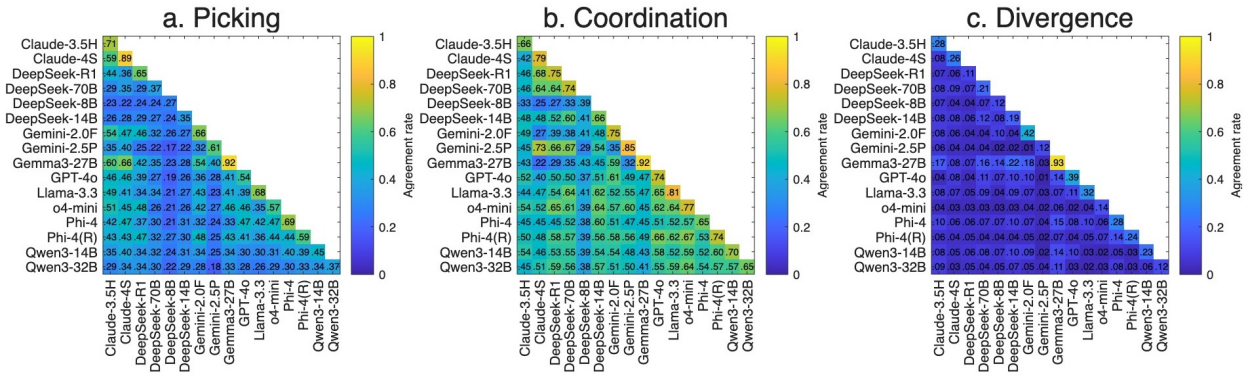
Notes: The table reports the average agreement rate and number of distinct answers across tasks for humans and LLMs subjects by treatment. For LLMs subjects we report the self-agreement rate. To ensure comparability between humans and LLM subjects (that were queried 50 times), we randomly subsample 50 human responses per treatment and calculate the average number of different responses.

Figure A.1 reports the average agreement rate among reasoning and non-reasoning models.[22] We find that reasoning models have lower agreement rates in the picking arm ($p < 0.05$ using Welch's t-test), that their performance in coordination tasks is comparable to non-reasoning models, and that they do better at divergence tasks ($p < 0.05$ using Welch's t-test).

To assess the role of LLM subjects facing exact copies of each other, we explore the performance of LLM subjects when facing a different LLM. Figure 4 shows the average (over all 12 topics) agreement rate for each pair of models by treatment condition (the diagonal corresponds to the self-agreement rates from Table 1). We find that the higher agreement rates among LLMs persist even when models are paired with different LLM subjects rather than an identical copy of themselves. The average agreement rate between a pair of LLMs is 37 percent in the picking treatment. This indicates that primary monoculture (salience) is still substantially higher than the human benchmark of 14 percent, even when the LLM subjects are not identical copies. The aver-

[22]Reasoning models break tasks into steps and write down their logic using a variety of approaches (e.g., chain of thought or tree of thought reasoning). These models excel in many domains, including math and logic (Li et al., 2025).

Figure 4: Pairwise agreement rates by treatment



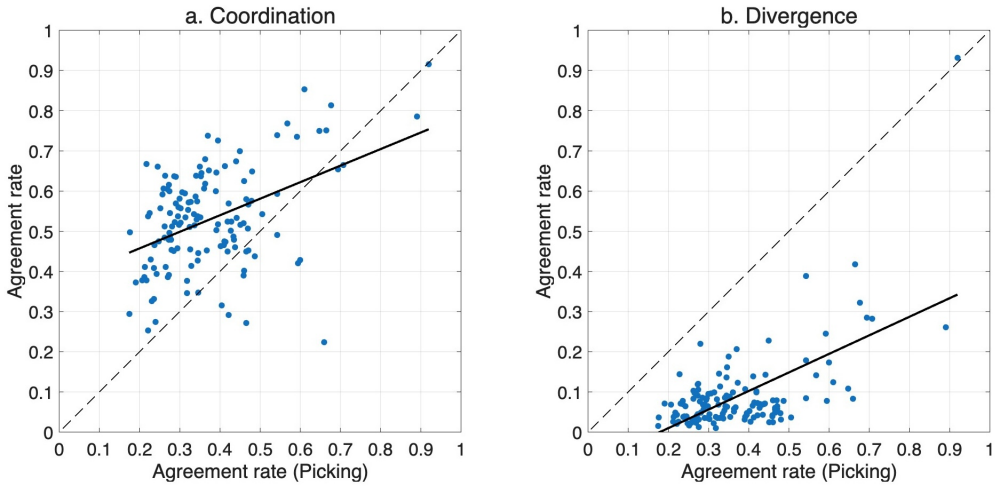
Notes: The figure displays a heatmap of pairwise agreement rates for all pairs of LLM subjects. Diagonal cells correspond to self-agreement rates (reported in Table 1). Lighter (yellow) colors indicate higher agreement and darker (blue) colors indicate lower agreement.

age agreement rate in the coordination treatment is 53 percent, still substantially higher than the 31 percent agreement rate obtained by human subjects. These findings point to coordination between different LLMs through strategic monoculture. Lastly, the agreement rate in the divergence arm is 9 percent, still substantially higher than the human benchmark, but substantially lower than the self-agreement rate. Consistent with our theoretical analysis, this finding points to large benefits from breaking symmetries in coordinated divergence games.

Finally, we evaluate the correlation between primary monoculture and performance in the coordination and divergence arms. Panel (a) of Figure 5 plots, for each pair of models, the agreement rate in the picking arm (on the x axis, corresponding to Panel (a) of Figure 4) against the agreement rate in the coordination arm (on the y axis, corresponding to Panel (b) of Figure 4). The fact that most points lie above the 45 degree line reflects the fact that the agreement rate increases in the coordination arm relative to the picking arm for almost any pair of LLMs. The figure also reveals a positive correlation between primary monoculture and the agreement rate in the coordination arm.

Panel (b) of Figure 5 repeats the analysis for the divergence arm. All but one pair of models lie under the 45-degree line (reflecting reduced agreement rates in the divergence arm). Furthermore, the figure reveals a positive correlation between primary monoculture and the agreement rate in the divergence arm. To summarize, Figure 5 shows that pairs of LLMs that provide similar answers in the absence of incentives tend to continue to provide similar answers both when incentivized to coordinate and when incentivized to diverge.

Figure 5: Primary monoculture and strategic coordination



Notes: Each point on a graph represents a pair of models. The x-axis reports their (average across topics) agreement rate in the picking arm, and the y-axis reports their (average across topics) agreement rates in the corresponding treatment arm. Panel (a) shows the coordination arm and panel (b) shows the divergence arm. Solid lines show the fitted OLS regression lines, and the dashed line indicates the 45° line.

5 Textual Analysis

In this section, we analyze the textual reasoning (Chain-of-Thought, or CoT for short) of LLM subjects.[23] We begin by analyzing the amount of LLM reasoning in the various treatment arms. To do this, we compare CoT length across treatment arms. In line with the difficulty to perform the task perfectly, we find that reasoning is lowest in the picking arm, and highest in the divergence arm (see Table A.4).[24]

Next, we turn to analyzing the semantic meaning of LLM reasoning.[25] We do so using two complementary methods: semantic similarity (as in Fish et al., 2024) and LLM as a judge (as in Asirvatham et al., 2026).

5.1 Semantic Similarity

Our semantic similarity analysis follows the methodology of Fish et al. (2024). We begin by splitting the LLM-generated reasoning text into individual sentences. This yields a dataset of 226,083 individual sentences. As we discuss above, in contrast to Fish et al. (2024), the amount of text is *not* balanced across treatment arms (for statistics on each model and treatment, see Table A.4).

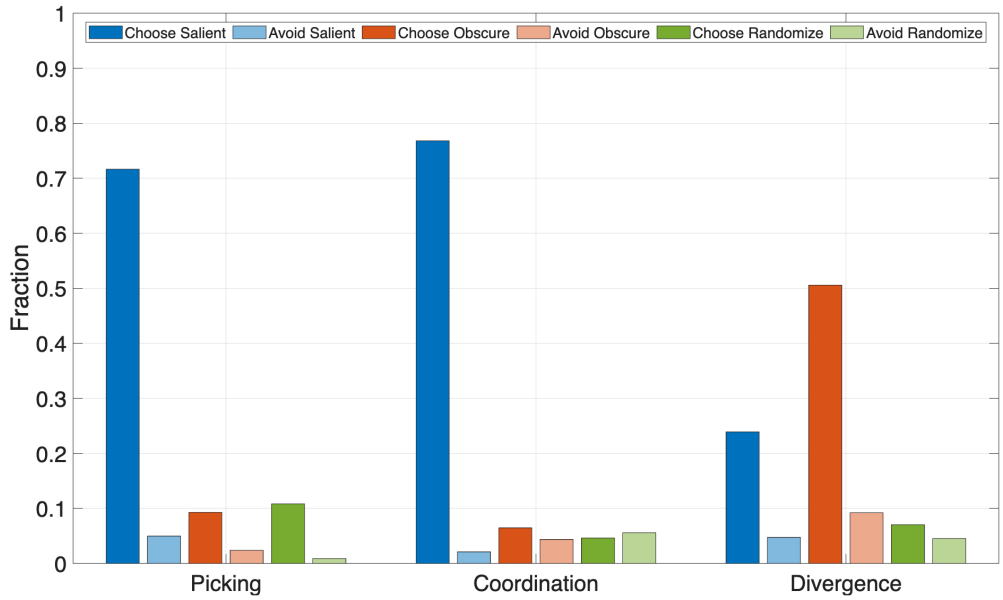
After splitting into sentences, we filter for sentences containing synonyms of the terms "salient," "obscure," and "randomize." (The full list of synonyms is available in Appendix E.1.) We note

[23] Some models only provide a summary of their reasoning. We omit these models from analyses that consider the length of CoT, and otherwise we analyze the summary of the reasoning.

[24] Interestingly, human subjects in the divergence arm also take more time to complete the experiment (see Table A.1).

[25] Fish et al. (2024) establish a causal link between LLM "thoughts" and actions in their setting.

Figure 6: Distribution of screened sentences by treatment



Notes: The figure reports the share of screened sentences that are semantically closest (in cosine similarity) to one of six reference vectors: CHOOSESALIENT, AVOIDSALIENT, CHOOSEOBSCURE, AVOIDOBSCURE, CHOOSERANDOMIZE, *and* AVOIDRANDOMIZE.

that the mere mention of a term like "salient" does not indicate an intention to choose a salient action. For example, the agent may be planning to *avoid* the salient choice. To address this issue, we incorporate the semantic meaning of each sentence into the analysis.

Following Fish et al. (2024), we analyze the semantic meaning of sentences using standard semantic similarity methods (e.g., Kenter and de Rijke, 2015; Mikolov et al., 2013). Specifically, we first use OpenAI's text-embedding-3-large to convert each screened sentence into a 3,072-dimensional vector. Next, for each term X of the three terms we screened for, we create two reference sentences, CHOOSEX and AVOIDX.[26] Given a screened sentence, we classify it according to the closest (in the sense of cosine similarity) reference vector.

Figure 6 summarizes our findings. The figure reveals that an overwhelming fraction of the sentences in the picking arm discuss salience favorably. The numbers are even higher in the coordination arm, consistently with our findings on models' behavior in Section 4. By contrast, most sentences in the divergence arm deal with obscurity. Furthermore, non-salience is emphasized relative to the coordination arms. Figure B.4 presents disaggregated results by model. These findings suggest that LLMs' reasoning is broadly consistent with the strategic environment of each treatment.

[26]Each reference vector is created by taking the average of the 3,072 dimensional vector corresponding to six sentences representative of the category. The exact sentences considered for each category are listed in Appendix E.1.

5.2 LLM as a Judge

This section presents a complementary textual analysis using LLM as a judge. This approach has several advantages relative to the semantic similarity analysis. First, it considers the entire text rather than screened sentences. Second, it enables us to consider the text as a whole, rather than on individual sentences. Third, it allows a more refined analysis of the semantic meaning of LLM-generated chains of thought.

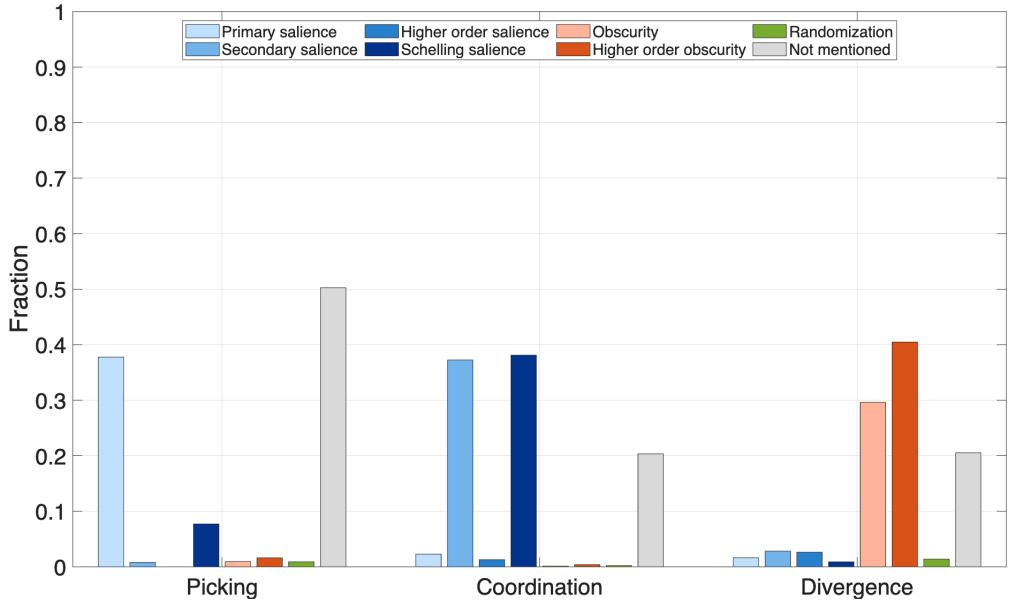
In the LLM-as-a-judge analysis, we feed the LLM-subjects' textual reasoning from the experiment to another LLM that serves as a "judge" (Gu et al., 2024, review this approach and its applications). The LLM judge is instructed to extract the main strategic reasoning expressed in the text as well as the perceived identity of the co-player (human or AI). The judge is informed about the general structure of the experiment and the topic of the question at hand, but is not given information about the experimental arm the specific CoT came from or about the LLM that generated the response.

Operationally, for each reasoning text, we use an independent call to our judge, which is implemented using Gemini 2.5 Flash. The judge is provided with 8 categories of strategic reasoning to choose from: primary salience, secondary salience, high-order salience, Schelling salience, obscurity, high-order obscurity, randomization, and "not mentioned." For each category, the judge is provided with a definition and examples. Additionally, it is provided with 8 choices for beliefs about the identity of the co-player (various degrees of certainty about the humanity of the co-player, as well as the possibility of it being an exact copy).[27] The full judge prompt is available in Appendix E.2.

Figure 7 summarizes the LLM-judge's determinations about LLM subjects' strategic reasoning. Our findings are in line with our results on LLM subjects' choices. First, in the picking arm, the most common strategic categories are "not mentioned" and primary salience. This reflects the lack of incentive to coordinate and points to a potential explanation for the high levels of primary monoculture among LLM subjects. Second, in the coordination arm, the most common strategic categories are secondary salience and Schelling salience. This provides further evidence of strategic algorithmic monoculture. Importantly, the increase in the strategic reasoning categories that reflect strategic algorithmic monoculture (68 percent) is substantially higher than the increase in agreement rate we can detect directly from choice data (14 percent). Third, in the divergence arm, the most common strategic categories are obscurity and higher-order obscurity. Again, this reflects strategic algorithmic monoculture (in the sense that agreement rates are regulated strategically), and the change in strategic reasoning is more drastic than the change we measure in choice data (68 percent vs. 31 percent). Figure B.4 presents the results for each model separately.

[27] This question is especially relevant for one of the experimental variations presented in the next section.

Figure 7: LLM-judge classification of strategic reasoning by treatment



Notes: The figure reports the share of LLM-reasoning texts that fall under each of the eight strategic reasoning categories according to the LLM-judge's determinations. The judge is implemented using Gemini 2.5 Flash. For details, see Appendix E.2.

6 Additional Variations

To probe the mechanisms behind our findings, we conducted several variations of the experiment. This section collects our findings.

6.1 Capabilities

A possible explanation for the high agreement rates among LLM subjects is that they only have access to a small number of valid answers. In the context of the divergence arm, it is also possible that they have limited ability to randomize between answers (Van Koevering and Kleinberg, 2024). In this section, we assess these capabilities.

To do so, we instruct each of the LLMs we study to generate a list of 100 valid answers (or, when there are fewer answers, as many as possible) . After completing this task, we ask each LLM to choose one answer from the list uniformly at random (the exact prompts are provided in Appendix C.4). We repeat this process 50 times for each LLM. We refer to the resulting choices as the *random* arm.

Table B.2 provides statistics on the number of valid answers each model generated. It reveals that models can readily produce many valid answers, ruling out lack of access as an explanation.

Next, we conduct two complementary analyses using the choices in the random arm. First, we calculate the agreement rate between LLM subjects in the random arm and in the divergence arm. We find that agreement rates are substantially lower than those achieved by LLM subjects in the divergence arm: on average, 3.83% when using the same model and 2.52% with a random

pair of models (Panel (a) of Figure B.5 displays the agreement rates for each pair of models). These findings indicate that LLM subjects are technically capable of achieving better outcomes in the divergence arm, suggesting that their poor performance results from other reasons. On the other hand, in the second analysis, we calculate the agreement rate between pairs of LLM subjects in the random arm (Panel b of Figure B.5). The average self-agreement rate is 12.20%, and the average agreement rate across models is 4.02%. This indicates that LLM subjects are not successful at following the instruction to randomize uniformly, and that part of the improvement in performance in the first analysis results from breaking the symmetry between LLM subjects.

6.2 Temperature

Since subjects are symmetric, our theoretical analysis suggests that success in the divergence arm crucially relies on the ability to randomize. On the other hand, randomization hinders coordination. To further investigate the role of the ability to randomize, we replicate the experiment using the highest and lowest temperatures allowed by each model (recall that the temperature is a parameter that modulates the randomness of LLM-generated text. Higher temperature corresponds to a higher degree of stochasticity in token prediction).[28]

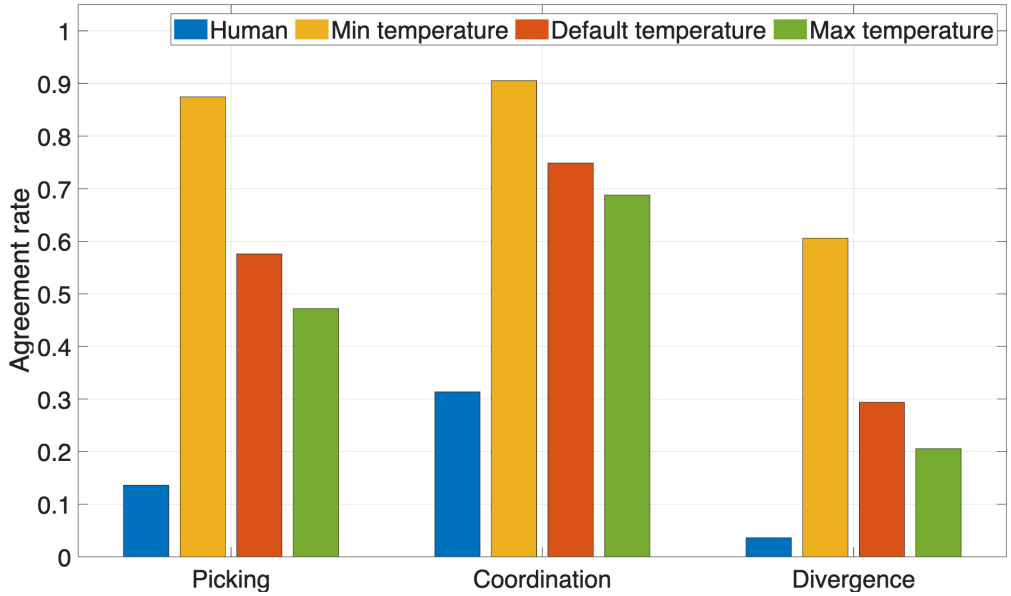
Figure 8 presents the aggregated results. Consistent with our theoretical analysis, when using the lowest temperatures, the agreement rate of these symmetric algorithms increases across all treatments ($p < 0.05$ for each pair, using Welch's t-test). When using the highest temperatures, agreement rates are lower across treatments, but the differences are not statistically significant. Finally, we note that even at these extreme temperatures, human subjects continue to outperform LLM subjects in the divergence arm, and the opposite holds in the coordination arm. Figure B.6 shows the results of Figure 8 for each LLM separately. The results continue to hold qualitatively. Altogether, our findings point to a tradeoff for users: increasing the temperature improves LLM subjects' performance in coordinated divergence tasks, but also hurts their performance in coordination tasks.

6.3 Information

Coordination through secondary salience and Schelling salience requires reasoning about the other player. In this section, we examine whether changing the information that is provided to LLM subjects about the identity of the other player affects coordination. In the main experiment (Section 3), LLM subjects were informed that the other subject was another LLM. In this section, we change the prompt so that the LLM is informed that (1) it is facing an identical *copy* of itself (the *copy* condition), or (2) that it is facing *"another person"* (the *person* condition). The exact textual instructions for these two conditions are provided in Figures C.2 and C.3.

[28] Our analysis covers 10 models, as we exclude models that provided unintelligible outputs under extreme temperatures, models that did not allow extreme temperature settings, and one model that was deprecated by the time we conducted this analysis.

Figure 8: Agreement rate under different temperature values by treatment



Notes: Bars represent the average self-agreement rates across topics for LLM subjects under different temperatures, for each treatment arm. For comparability, the default temperature average is calculated only over models that were not excluded, and human agreement rates are also displayed. Differences between default and low temperatures are statistically significant across all treatment arms ($p < 0.05$), whereas differences between default and high temperatures are not.

To quantify the effect of providing this information, we estimate the following linear regression (separately for each condition):

$$A_{it} = \alpha + \beta\, T_{it} + \mu_i + \delta_t + \varepsilon_{it}, \tag{2}$$

where the variable A_{it} is the self-agreement rate of model i in topic t and T_{it} is an indicator equal to one if the model is assigned to the treatment condition (either copy or person), and zero otherwise. The specification also includes model fixed effects, μ_i, and topic fixed effects, δ_t. The coefficient of interest is β, which measures the causal effect of changing the information about the other player's identity on strategic behavior.

Table 2 displays the estimated β coefficients with 95% confidence intervals for both the copy and the person conditions. LLM performance improved in both treatment arms under the copy condition. Specifically, agreement rates increase in the coordination arm and decrease in the divergence arm for most LLMs. However, these effects are small and not statistically significant. Similarly, we find no statistically significant effects in the person condition.

The average effects reported above mask a substantial heterogeneity between reasoning and non-reasoning models. Figure A.2 reveals that non-reasoning models (whose performance is weaker at the baseline condition) benefit substantially (and significantly) in the divergence arm from learning that they are playing a copy. Figure B.7 shows the effect by model.

Table 2: Effect of information treatment on self-agreement rates

	(1) Coordination	(2) Divergence
Copy	0.017	−0.020
	(0.014)	(0.024)
Person	0.009	0.012
	(0.011)	(0.015)

Notes: The table presents the results of the linear regression specified in Equation (2), which includes model and topic fixed effects. The dependent variable is the self-agreement rate in the coordination and divergence arms, respectively. *Copy* and *Person* are indicators that equal one in the copy and person conditions, respectively. Standard errors in parentheses are computed using the CR2 variance estimator and clustered by model. * p<0.10, ** p<0.05, *** p<0.01.

Textual analysis. We repeat the textual analyses from Section 5. We find that the length of reasoning in the divergence arm under the copy condition is even higher than in the main divergence arm (see Table A.4). Additionally, relative to the divergence arm of the main experiment, our semantic similarity analysis finds that in the copy condition, discussion of randomization becomes more prevalent while the overall level of obscurity decreases slightly. By contrast, the person condition has little effect across experimental arms (see Figure A.3). Similarly, the LLM-as-a-Judge analysis also documents an increase in the randomization category under the copy condition in the divergence arm. It goes beyond the semantic similarity analysis, by documenting a sharp increase in the Schelling salience category under the copy condition in the coordination arm (mostly replacing secondary salience reasoning).

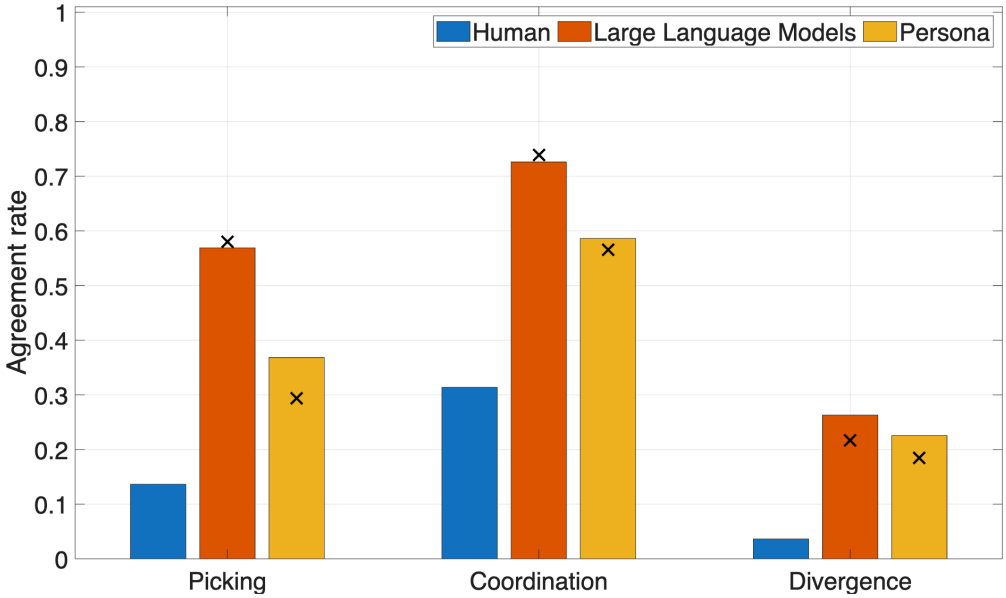
Recall that the LLM judge was also instructed to classify the reasoning with respect to beliefs about the identity of the other player. We find that the judge's classification is consistent with the expected effect of our experimental conditions. Specifically, across all conditions, the results are nearly identical in the divergence in coordination arms. In the baseline experiment, the most common categories are LLM co-player and "not mentioned," in the copy condition, the most common responses are copy and "not mentioned," and in the person condition the most common responses are Human and "not mentioned" (see Figure A.4 for aggregated results and Figure B.10 for results disaggregated by model).

6.4 Personas

In this section, we consider an established approach for increasing heterogeneity: personas. As in Mei et al. (2024), we replicate the main experiment informing LLM subjects that they are playing for a human with specific characteristics.[29] Specifically, to mimic the distribution of traits in the human-subject sample, we randomly draw 50 different human subjects and construct personas based on their gender, age, occupation, race/ethnicity, and education level. Operationally, we provide the persona information first (see Appendix C.3) and then use the prompt from the person

[29] We thank Alex Imas for proposing this analysis. We use only 14 models, as two models had been deprecated by the time we conducted this analysis.

Figure 9: Agreement rate by LLMs with personas



Note: Bars report average agreement rate across question topics for human and LLM subjects by treatment. LLM agreement rates are the average over all LLMs when facing the same model. The black cross-shaped markers indicate the median.

condition.[30] We use the same 50 personas for all topics and treatment arms, but rely on a separate, independent LLM calls (yielding a total of $3 \times 50 \times 12 = 1,800$ independent calls per model).

Figure 9 summarizes the results. Although personas shift LLM subjects' behavior closer to human subjects, our main results remain qualitatively unchanged. In particular, even when endowed with heterogeneous personas, LLM subjects significantly lag behind humans in the divergence arm. Figures A.5, A.6 and B.11 show similar results in replications of other analyses from Section 4.

7 Discussion

The role of AI agents in society is rapidly increasing. These agents do not operate in a vacuum. Rather, they interact with each other, forming complex multi-agent systems. These systems pose novel risks, many of which are linked to undesirable coordination (collusion) and miscoordination (Hammond et al., 2025).

Against this backdrop, we make two contributions. Conceptually, we introduce the taxonomy of primary versus strategic algorithmic monoculture, paralleling the distinction of the human-coordination literature between primary salience and other forms of salience (Mehta et al., 1994; Schelling, 1960). Empirically, we implement a simple experimental design that allows us to separate *strategic* monoculture from primary monoculture.

[30] We use the person condition prompt because the main analysis prompt says that the co-player is "another LLM" which contradicts the persona instructions. For the picking arm, we use the main analysis prompt, as there is no such contradiction.

The results of our experiment reveal a sharp asymmetry. With no incentives to agree or disagree, LLM choices are substantially more homogeneous than human choices, consistent with elevated primary monoculture. And while both humans and LLMs respond to incentives—increasing or decreasing agreement when it is rewarded—their performance varies. LLMs dramatically outperform humans in coordination on the same action, but humans dramatically outperform LLMs in divergence. In short, LLMs are exceptionally effective at converging on focal responses, but comparatively poor at sustaining socially valuable heterogeneity.

Our findings contribute to the discussion of monoculture, model multiplicity, and related multi-agent risks (Bommasani et al., 2021; Gorecki and Hardt, 2025; Raghavan, 2025). The concerns expressed in this literature manifest, for example, when admissions or hiring decisions are delegated to algorithms, or assisted by them (Kleinberg and Raghavan, 2021). Intuitively, as the number of models increases, even if their errors are partially correlated, one can hope that college applicants or job candidates will have *recourse*—i.e., that they will not be systematically excluded due to noise. We contribute to this literature by showing that LLMs regulate the similarity of their choices in response to strategic incentives. In richer environments—such as hiring or admissions—strategic incentives (including avoiding adverse selection or congestion, Ali et al., 2026; Baek et al., 2026) may further shape the correlation between various algorithms, a consideration that is often overlooked. We hope that our perspective on algorithmic monoculture will advance our understanding of these important risks and help in identifying ways to mitigate them.

References

Akata, E. et al. (2025). Playing Repeated Games with Large Language Models. *Nature Human Behaviour*, pages 1–11.

Ali, S. N., Candelas, S., and Shorrer, R. I. (2026). When Colleges Compete: Signals, Noise, and Equilibrium Outcomes. Technical report, Working paper, Penn State.

Araujo, D. K. and Uhlig, H. (2026). How Does AI Distribute the Pie? Large Language Models and the Ultimatum Game. Working Paper 34919, National Bureau of Economic Research.

Asirvatham, H., Mokski, E., and Shleifer, A. (2026). GPT as a Measurement Tool. Technical report, National Bureau of Economic Research.

Assad, S. et al. (2024). Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market. *Journal of Political Economy*, 132(3):723–771.

Baek, J., Bastani, H., and Chen, S. (2026). Strategic hiring under algorithmic monoculture.

Bardsley, N., Mehta, J., Starmer, C., and Sugden, R. (2010). Explaining Focal Points: Cognitive Hierarchy Theory Versus Team Reasoning. *The Economic Journal*, 120(543):40–79.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*.

Brynjolfsson, E., Korinek, A., and Agrawal, A. (2024). The Economics of Transformative AI: A Research Agenda. Unpublished manuscript.

Calvano, E. et al. (2020a). Artificial Intelligence, Algorithmic Pricing, and Collusion. *American Economic Review*, 110(10):3267–3297.

Calvano, E. et al. (2020b). Protecting Consumers from Collusive Prices Due to AI. *Science*, 370(6520):1040–1042.

Camerer, C. F., Ho, T.-H., and Chong, J.-K. (2004). A Cognitive Hierarchy Model of Games. *The Quarterly Journal of Economics*, 119(3):861–898.

Cao, C., Wang, K. W., and Yang, D. (2026). When AI Fails: A Natural Experiment on Generative AI's Role in Crypto Markets. Working paper, January 2026.

Charness, G., Jabarian, B., and List, J. A. (2023). Generation Next: Experimentation with AI. Working Paper 31679, National Bureau of Economic Research.

Chen, Y., Liu, T. X., Shan, Y., and Zhong, S. (2023). The Emergence of Economic Rationality of GPT. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120.

Cheng, Q., Lin, P., and Zhao, Y. (2025). Does Generative AI Facilitate Investor Trading? Early Evidence from ChatGPT Outages. *Journal of Accounting and Economics*, page 101821.

Conitzer, V. and Oesterheld, C. (2023). Foundations of cooperative ai. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*.

Cook, T. R., Kazinnik, S., Modig, Z., and Palmer, N. M. (2026). What Do LLMs Want? *Federal Reserve Working Paper No. 2026-006*.

Cooper, R. (1999). *Coordination Games*. Cambridge University Press.

Crawford, V. P., Costa-Gomes, M. A., and Iriberri, N. (2013). Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature*, 51(1):5–62.

Crawford, V. P. and Iriberri, N. (2007). Fatal attraction: Salience, naivete, and sophistication in experimental "hide-and-seek" games. *American Economic Review*, 97(5):1731–1750.

Dreyfuss, B. and Raux, R. (2024). Human Learning about AI Performance. *arXiv preprint arXiv:2406.05408*.

Eicher, J. E. and Irgolic, R. F. (2024). Compensatory Biases under Cognitive Load: Reducing Selection Bias in Large Language Models. *CoRR*, abs/2402.01740.

Ericsson, K. A. and Simon, H. A. (1998). How to Study Thinking in Everyday Life: Contrasting Think-Aloud Protocols with Descriptions and Explanations of Thinking. *Mind, Culture, and Activity*, 5(3):178–186.

Fish, S. et al. (2025). Econevals: Benchmarks and Litmus Tests for LLM Agents in Unknown Environments. *arXiv preprint arXiv:2503.18825*.

Fish, S., Gonczarowski, Y. A., and Shorrer, R. I. (2024). Algorithmic Collusion by Large Language Models. *arXiv preprint arXiv:2404.00806*, 7.

Gorecki, M. and Hardt, M. (2025). Monoculture or Multiplicity: Which is It? In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al. (2024). A Survey on LLM-as-a-Judge. *The Innovation*.

Guo, Z., Lv, H., Zhang, C., Zhao, Y., Zhang, Y., and Cui, L. (2025). The Illusion of Randomness: How LLMs Fail to Emulate Stochastic Decision-Making in Rock-Paper-Scissors Games? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8618–8637.

Hammond, L. et al. (2025). Multi-Agent Risks from Advanced AI. *arXiv preprint arXiv:2502.14143*.

Harsanyi, J. C. and Selten, R. (1988). A General Theory of Equilibrium Selection in Games. *MIT Press Books*, 1.

Hassan, T. A. and Mertens, T. M. (2011). Market Sentiment: A Tragedy of the Commons. *American Economic Review*, 101(3):402–05.

Hassan, T. A. and Mertens, T. M. (2017). The Social Cost of Near-Rational Investment. *American Economic Review*, 107(4):1059–1103.

He, K., Shorrer, R., and Xia, M. (2025). Human Misperception of Generative-AI Alignment: A Laboratory Experiment. *arXiv preprint arXiv:2502.14708*.

Heath, C., Ho, B., and Berger, J. (2006). Focal Points in Coordinated Divergence. *Journal of Economic Psychology*, 27(5):635–647.

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

Hosseini, H. and Khanna, S. (2025). Distributive Fairness in Large Language Models: Evaluating Alignment with Human Values. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*.

Imas, A., Lee, K., and Misra, S. (2025). Agentic Interactions. *Available at SSRN 5875162*.

Isoni, A., Poulsen, A., Sugden, R., and Tsutsui, K. (2014). Efficiency, Equality, and Labeling: An

Experimental Investigation of Focal Points in Explicit Bargaining. *American Economic Review*, 104(10):3256–3287.

Jackson, M. O. et al. (2025). AI Behavioral Science. *SSRN Electronic Journal*.

Jiang, L. et al. (2025). Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond). *CoRR*, abs/2510.22954.

Kenter, T. and de Rijke, M. (2015). Short Text Similarity with Word Embeddings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM)*, pages 1411–1420.

Kim, E. et al. (2025). Correlated Errors in Large Language Models. *CoRR*, abs/2506.07962.

Kleinberg, J. and Raghavan, M. (2021). Algorithmic Monoculture and Social Welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118.

Korinek, A. (2023). Generative AI for Economic Research: Use Cases and Implications for Economists. *Journal of Economic Literature*, 61(4):1281–1317.

Li, Z. et al. (2025). From System 1 to System 2: A Survey of Reasoning Large Language Models. *CoRR*, abs/2502.17419.

Manning, B. S., Zhu, K., and Horton, J. J. (2024). Automated Social Science: Language Models as Scientist and Subjects. arXiv:2404.11794 [econ].

Mehta, J., Starmer, C., and Sugden, R. (1994). The Nature of Salience: An Experimental Investigation of Pure Coordination Games. *The American Economic Review*, 84(3):658–673.

Mei, Q., Xie, Y., Yuan, W., and Jackson, M. O. (2024). A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121.

Meta FAIR Diplomacy Team et al. (2022). Human-Level Play in the Game of Diplomacy by Combining Language Models with Strategic Reasoning. *Science*, 378(6624):1067–1074.

Mikolov, T. et al. (2013). Efficient Estimation of Word Representations in Vector Space. In *International Conference on Learning Representations*.

Page, S. (2008). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press.

Peng, K. and Garg, N. (2024). Monoculture in Matching Markets. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Pezeshkpour, P. and Hruschka, E. (2024). Large Language Models Sensitivity to the Order of Options in Multiple-Choice Questions. In Duh, K., Gómez-Adorno, H., and Bethard, S., editors,

Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024, pages 2006–2017. Association for Computational Linguistics.

Raghavan, M. (2025). Competition and Diversity in Generative AI.

Rangel, A., Camerer, C., and Montague, P. R. (2008). A Framework for Studying the Neurobiology of Value-Based Decision Making. *Nature reviews neuroscience*, 9(7):545–556.

Rilla, R. et al. (2025). Recognising, Anticipating, and Mitigating LLM Pollution of Online Behavioural Research. *arXiv preprint arXiv:2508.01390*.

Ross, J., Kim, Y., and Lo, A. W. (2024). LLM Economicus? Mapping the Behavioral Biases of LLMs via Utility Theory. *arXiv preprint arXiv:2408.02784*.

Schelling, T. (1960). *The Strategy of Conflict*. Harvard University Press.

Society for Human Resource Management (2025). 2025 Talent Trends. Research report, Society for Human Resource Management.

Sugden, R. (1995). A Theory of Focal Points. *The Economic Journal*, 105(430):533–550.

Vafa, K., Rambachan, A., and Mullainathan, S. (2024). Do Large Language Models Perform the Way People Expect? Measuring the Human Generalization Function. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24.

Van Koevering, K. and Kleinberg, J. (2024). How Random is Random? Evaluating the Randomness and Humaness of LLMs' Coin Flips. *arXiv preprint arXiv:2406.00092*.

Wenger, E. and Kenett, Y. (2025). We're Different, We're the Same: Creative Homogeneity Across LLMs.

Werner, T. et al. (2024). Experimental Evidence That Conversational Artificial Intelligence Can Steer Consumer Behavior Without Detection.

Zhang, L. H., Milli, S., Jusko, K., Smith, J., Amos, B., Bouaziz, W., Revel, M., Kussman, J., Sheynin, Y., Titus, L., et al. (2025). Cultivating Pluralism in Algorithmic Monoculture: The Community Alignment Dataset. *arXiv preprint arXiv:2507.09650*.

Supplementary Materials

A Additional Tables and Figures

Table A.1: Summary statistics (human subjects)

Variable	Picking	Coordination	Divergence
Age	47.47	47.09	45.33
	(14.03)	(11.91)	(12.63)
Female	0.50	0.54	0.57
	(0.50)	(0.50)	(0.50)
Duration (in minutes)	3.91	3.84	4.92
	(0.23)	(0.20)	(0.21)
Tab switches	0.50	0.40	0.55
	(0.62)	(0.62)	(0.58)
Number of Observations	105	99	97

Notes: Values are means with standard deviations in parentheses.

Table A.2: Features of LLM subjects

Type	**Provider**	**Label**	**Model**	**Reasoning**	**Deafult Temperature**
Closed source	Anthropic	Claude-3.5H	Claude-3.5-Haiku	No	1
		Claude-4S	Claude-4-Sonnet	Yes	1
	Google	Gemini-2.0F	Gemini-2.0-Flash	No	1
		Gemini-2.5P	Gemini-2.5-Pro	Yes	1
	OpenAI	GPT-4o	GPT-4o	No	1
		o4-mini	o4-mini	Yes	1
Open source	DeepSeek	DeepSeek-8B	DS-Llama-8B	Yes	0.5
		DeepSeek-70B	DS-Llama-70B	Yes	0.5
		DeepSeek-14B	DS-Qwen-14B	Yes	0.5
		DeepSeek-R1	DS-R1-671B	Yes	1
	Google	Gemma3-27B	Gemma3-27B	Yes	0.5
	Llama	Llama-3.3	Llama-3.3-70B	No	1
	Microsoft	Phi-4	Phi-4-14B	No	0.5
		Phi-4(R)	Phi-4-Reasoning	Yes	0.5
	Qwen	Qwen3-32B	Qwen3-32B	Yes	0.5
		Qwen3-14B	Qwen3-14B	Yes	0.5

Notes: The table provides information on LLM subjects. In particular, it details the full name of the model alongside the label used in the paper, and whether the model has reasoning abilities.

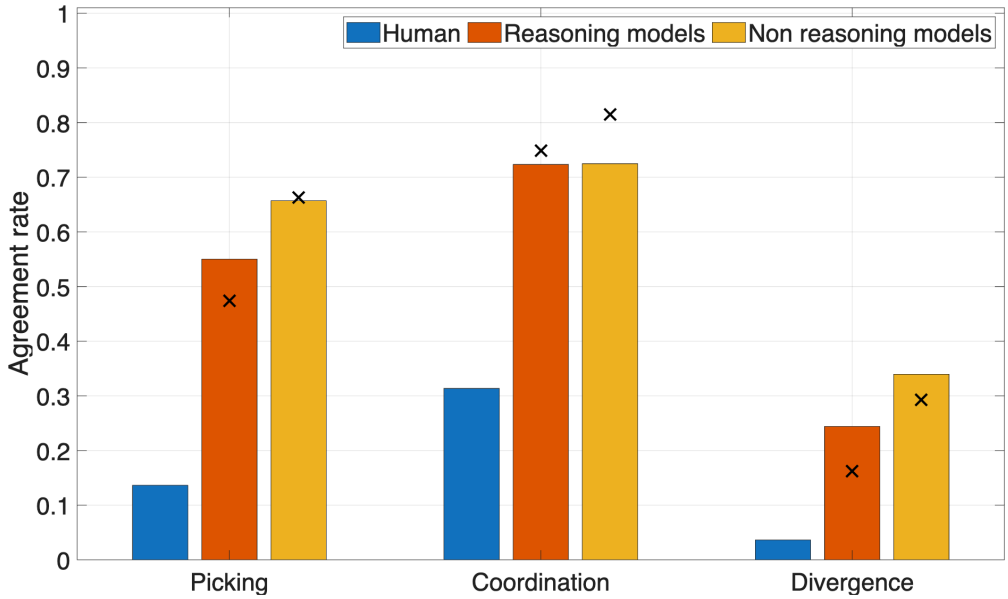
Table A.3: Most frequent responses by treatment, topic, and subject type

	Picking				**Coordination**				**Divergence**			
	Human		LLMs		Human		LLMs		Human		LLMs	
Topic	Response	Freq.	Response	Freq.	Response	Freq.	Response	Freq.	Response	Freq.	Response	Freq.
athlete	michael jordan	25.00	lebron james	67.60	michael jordan	28.30	lebron james	37.00	lebron james	7.40	lebron james	13.40
	lebron james	19.20	lionel messi	13.90	lebron james	21.20	lionel messi	33.50	michael jordan	4.30	shohei ohtani	7.60
	lionel messi	3.80	serena williams	7.10	tom brady	8.10	cristiano ronaldo	13.80	simone biles	4.30	simone biles	7.40
	serena williams	3.80	cristiano ronaldo	4.00	travis kelce	7.10	michael jordan	13.00	usain bolt	4.30	serena williams	5.80
	dak prescott	2.90	michael jordan	2.40	kobe bryant	6.10	serena williams	0.50	aaron judge	3.20	mikaela shiffrin	4.40
cars	ford	28.60	toyota	92.10	ford	47.00	toyota	97.80	ford	11.60	volvo	14.40
	toyota	22.90	ford	5.00	toyota	23.00	ford	2.10	honda	5.30	mazda	13.90
	honda	12.40	tesla	1.50	honda	13.00	honda	0.10	hyundai	5.30	subaru	13.90
	chevrolet	4.80	bmw	0.60	chevrolet	8.00	abarth	0.00	kia	5.30	toyota	8.40
	kia	4.80	honda	0.30	nissan	2.00	acura	0.00	nissan	5.30	tesla	5.30
city	paris	14.40	tokyo	44.70	new york	50.00	paris	53.30	oslo	3.30	kyoto	16.60
	new york	10.60	paris	39.20	paris	23.00	new york	27.10	barcelona	2.20	ulaanbaatar	16.60
	london	5.80	kyoto	6.90	london	4.00	london	9.50	berlin	2.20	reykjavik	7.00
	chicago	3.80	london	3.30	los angeles	4.00	tokyo	7.70	dallas	2.20	valparaiso	2.40
	houston	3.80	sydney	2.90	chicago	2.00	istanbul	0.60	dubai	2.20	tokyo	2.20
color	blue	43.80	blue	77.50	blue	60.00	blue	81.50	purple	11.50	teal	26.50
	red	17.10	cerulean	7.40	red	29.00	red	17.60	orange	8.30	chartreuse	13.70
	green	11.40	red	4.10	green	3.00	green	0.60	magenta	7.30	purple	8.50
	purple	7.60	teal	2.20	purple	3.00	magenta	0.10	pink	7.30	turquoise	8.20
	black	5.70	sky blue	1.60	black	2.00	purple	0.10	red	7.30	cerulean	8.00
disease	cancer	32.70	diabetes	29.20	cancer	51.00	influenza	32.90	cancer	9.50	whipple's	7.80
	aids	7.70	influenza	21.50	covid	15.00	diabetes	18.30	lupus	7.40	tuberculosis	7.20
	diabetes	5.80	alzheimer's	20.30	aids	5.00	cancer	15.70	crohn's disease	6.30	kawasaki	6.60
	influenza	5.80	malaria	10.00	influenza	5.00	covid-19	14.70	diabetes	5.30	creutzfeldt-jakob	6.00
	measles	4.80	tuberculosis	7.30	heart disease	4.00	alzheimer's	11.40	lyme disease	4.20	kuru	5.00
flower	rose	48.60	rose	70.90	rose	84.00	rose	93.50	rose	9.50	sunflower	10.40
	daisy	20.00	sunflower	12.60	daisy	8.00	sunflower	3.80	daisy	7.40	orchid	10.20
	lily	6.70	daisy	5.20	tulip	3.00	daisy	1.10	tulip	7.40	jasmine	6.70
	iris	4.80	tulip	4.40	lily	2.00	magnolia	0.50	daffodil	6.30	bleeding heart	3.90
	tulip	3.80	lily	2.60	sunflower	2.00	hydrangea	0.40	sunflower	6.30	dahlia	3.90
food	pizza	23.80	pizza	52.90	pizza	43.00	pizza	74.50	pizza	7.30	sushi	13.50
	burger	8.60	sushi	14.40	burger	9.00	apple	17.10	potato	4.20	quinoa	6.60
	bread	7.60	apple	12.90	apple	8.00	bread	5.70	egg	3.10	pizza	5.00
	apple	5.70	spaghetti	2.80	bread	8.00	rice	0.90	apple	2.10	pineapple	3.60
	banana	3.80	avocado	2.70	chicken	3.00	avocado toast	0.30	asparagus	2.10	kimchi	3.00
geometric	triangle	32.70	triangle	38.50	square	49.00	circle	62.70	triangle	15.80	pentagon	19.90
	square	26.90	circle	32.40	circle	23.00	square	20.60	rhombus	12.60	hexagon	11.10
	circle	17.30	square	11.20	triangle	22.00	triangle	12.90	octagon	10.50	heptagon	9.50
	hexagon	5.80	hexagon	5.40	hexagon	2.00	cube	2.60	hexagon	9.50	dodecahedron	7.60
	octagon	5.80	cube	4.00	cube	1.00	hexagon	1.00	square	5.30	trapezoid	6.50
letter	a	41.90	a	27.30	a	73.00	e	50.10	a	7.30	q	32.70
	e	9.50	m	16.30	e	6.00	a	46.30	j	6.30	z	12.70
	b	8.60	q	9.70	b	3.00	m	2.60	w	6.30	k	11.10
	z	4.80	g	9.20	c	3.00	b	0.40	z	6.30	a	9.90
	d	3.80	e	7.80	s	3.00	c	0.30	l	5.20	x	7.60
month	april	12.40	january	22.80	november	36.00	january	51.00	february	16.70	september	29.80
	may	11.40	october	17.00	january	29.00	july	15.00	march	13.50	february	18.70
	november	11.40	march	16.90	december	6.00	december	8.50	august	11.50	april	9.80
	january	10.50	july	14.20	june	6.00	october	8.50	june	10.40	august	7.40
	march	10.50	april	13.10	july	5.00	june	7.10	april	9.40	november	6.90
positive number	2	20.00	7	37.20	2	33.00	1	78.80	8	12.50	7	14.80
	4	11.40	42	32.80	1	22.00	7	13.70	12	5.20	1	10.70
	5	10.50	5	11.80	7	11.00	42	5.30	2	4.20	2	9.00
	1	8.60	1	5.80	4	7.00	5	1.10	7	4.20	17	8.10
	8	7.60	17	2.60	10	5.00	17	0.50	22	3.10	42	7.20
trait	honesty	18.10	empathy	28.40	honesty	31.60	honesty	56.70	honesty	20.20	empathy	19.90
	kindness	10.50	honesty	15.60	kindness	11.20	kindness	20.20	humility	7.40	resilience	19.80
	empathy	5.70	compassion	13.60	funny	8.20	empathy	12.50	kindness	4.30	perseverance	6.40
	funny	5.70	kindness	10.20	friendly	4.10	loyalty	2.30	loyal	4.30	meticulous	4.70
	friendly	4.80	resilience	9.00	smart	4.10	resilience	2.10	shy	4.30	curiosity	4.30

Table A.4: Number of reasoning sentences by model and treatment

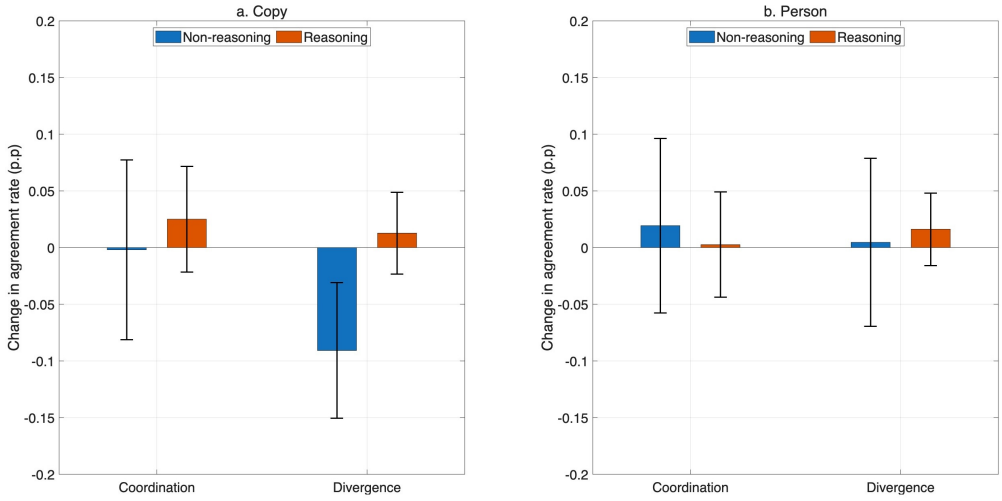
Model	Picking	Coordination	Divergence	Coordination (Copy)	Divergence (Copy)	Coordination (Person)	Divergence (Person)	Total
Claude-3.5H	0	3,080	2,199	2,562	1,949	1,948	1,534	13,272
Claude-4S	0	4,017	4,140	3,922	5,685	3,752	3,938	25,454
DeepSeek-14B	12,206	26,689	51,691	34,635	67,941	28,964	52,977	275,103
DeepSeek-70B	9,945	19,311	48,469	26,449	58,785	22,795	44,406	230,160
DeepSeek-8B	11,840	22,776	31,985	31,288	53,601	24,044	35,806	211,340
DeepSeek-R1	6,884	40,752	107,375	56,113	176,669	36,530	107,968	532,291
Gemini-2.0F	0	247	265	1,539	1,601	368	257	4,277
Gemini-2.5P	12,918	20,798	19,466	19,368	19,616	19,853	19,723	131,742
Gemma3-27B	304	129	119	105	117	149	274	1,197
GPT-4o	1	0	1	0	3	2	0	7
Llama-3.3	0	1,993	1,815	2,928	3,826	1,535	1,403	13,500
o4-mini	10	2,110	2,476	2,187	2,956	2,285	2,646	14,670
Phi-4	5	2,472	1,755	2,418	3,750	4,205	3,246	17,851
Phi-4(R)	89,822	130,431	205,138	135,132	251,805	140,195	189,997	1,142,520
Qwen3-14B	9,318	38,859	68,919	52,553	95,254	42,774	69,758	377,435
Qwen3-32B	10,437	35,728	75,511	41,016	101,187	37,836	71,130	372,845
Total	163,690	349,392	621,324	412,215	844,745	367,235	605,063	3,363,664

Figure A.1: Agreement rate by treatment and model capabilities



Notes: Bars report the average self-agreement rate across topics by treatment and reasoning abilities. LLM self-agreement rates are the average over all LLMs with the same reasoning capability. For comparability, we also report human agreement rates. The black cross-shaped markers indicate the median.

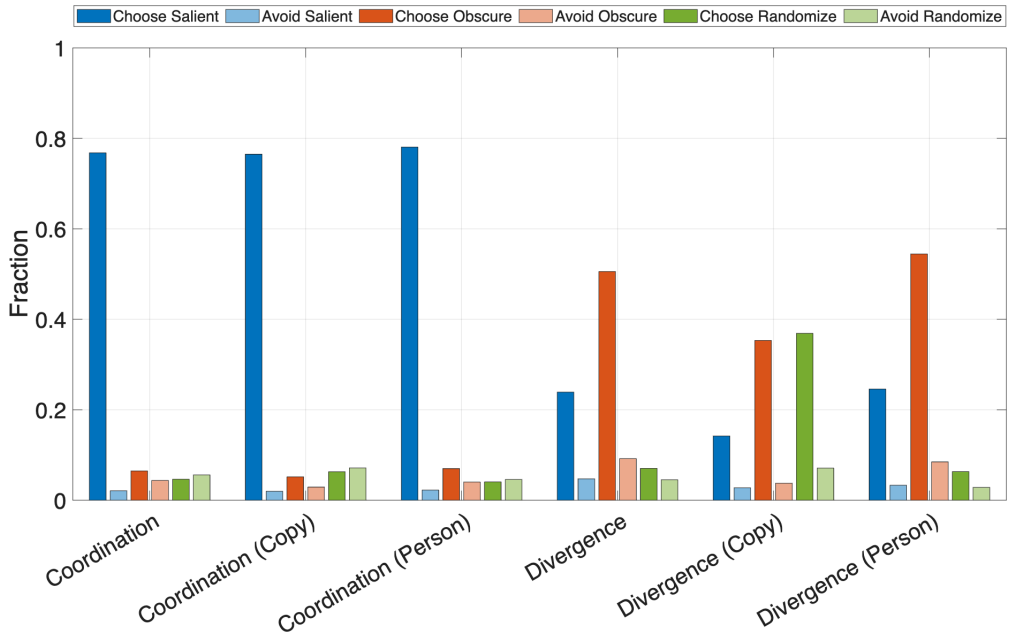
Figure A.2: Effect of information on agreement rate by model capabilities and condition



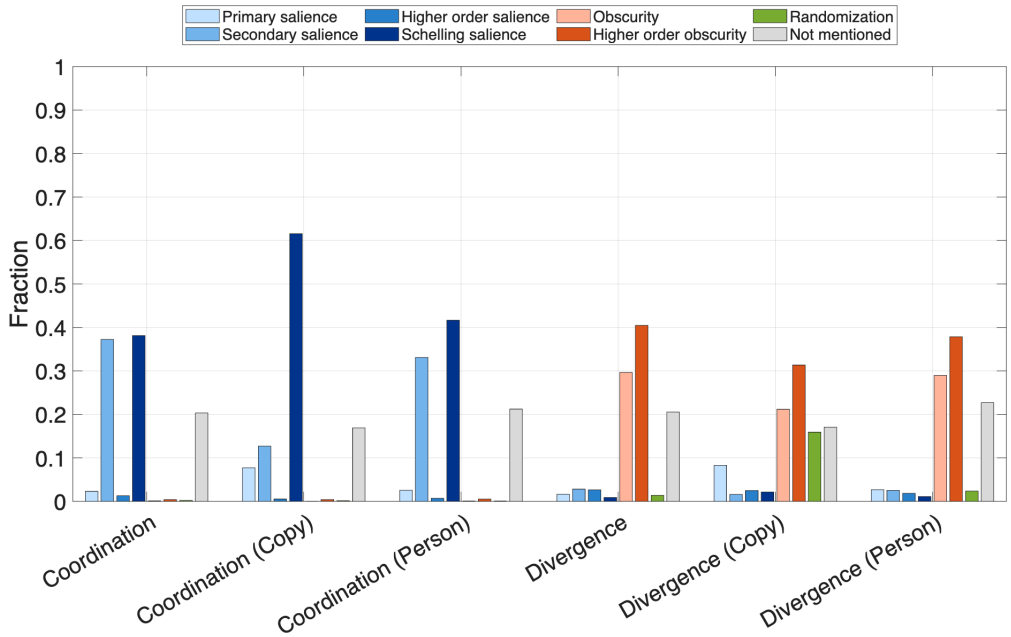
Notes: Bars represent the average absolute change in self-agreement rates across topics for LLM subjects by treatment and reasoning capabilities. Panel (a) corresponds to the copy condition and Panel (b) to the person condition. Vertical lines indicate 95% confidence intervals.

Figure A.3: Textual analysis of strategic reasoning in the copy and person condition

(a) Distribution of screened sentences

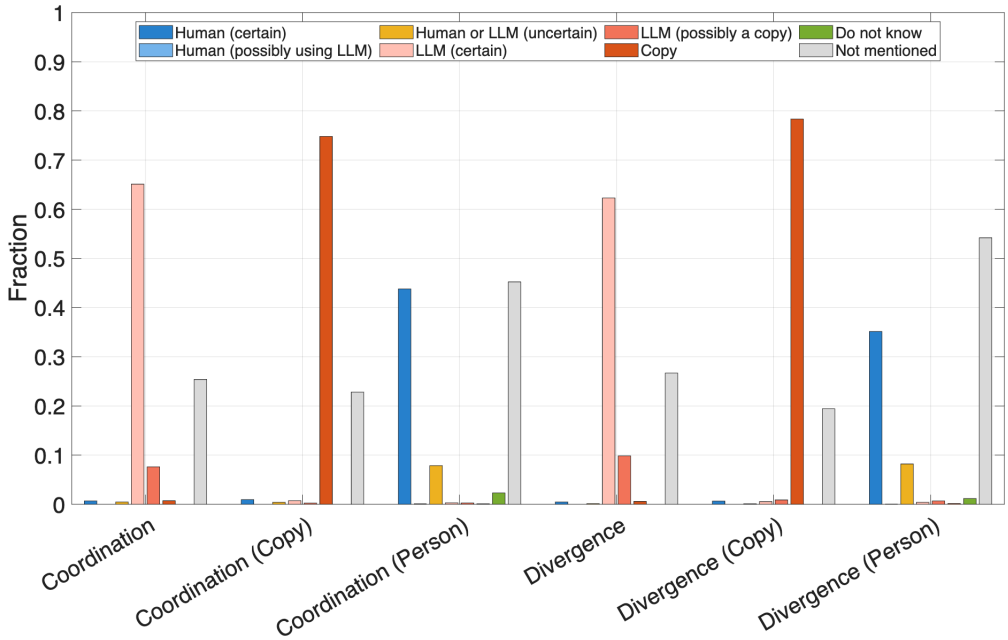


(b) LLM-judge classification of strategic reasoning



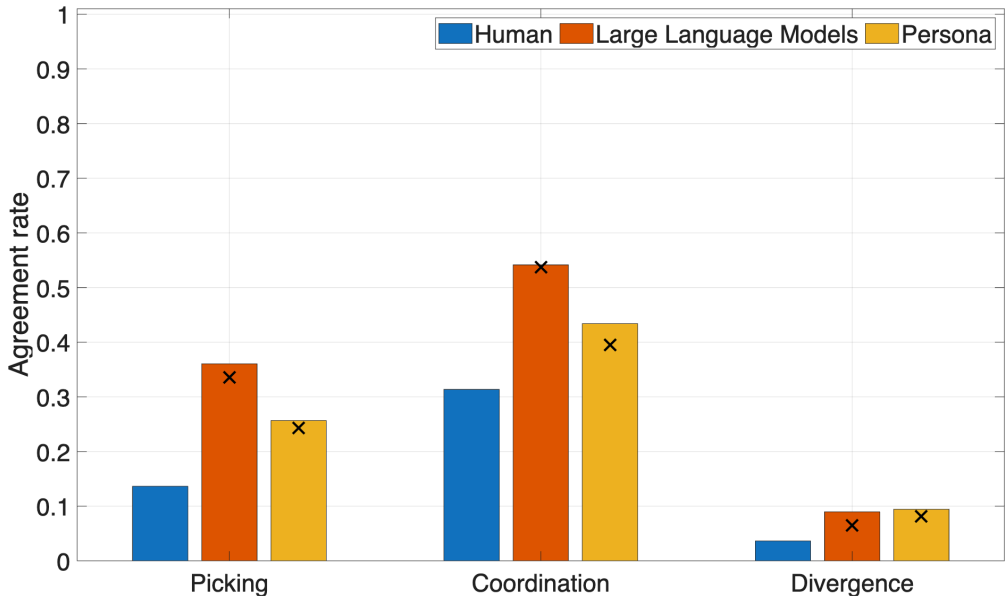
Notes: Panel (a) displays the distribution of screened sentences' closest reference vector by treatment. Panel (b) reports the share of LLM-reasoning texts that fall under each of the eight strategic reasoning categories according to the LLM-judge's determination. The judge is implemented using Gemini 2.5 Flash. For details, see Appendix E.2.

Figure A.4: LLM-judge classification of beliefs about co-player identity (by treatment and condition)



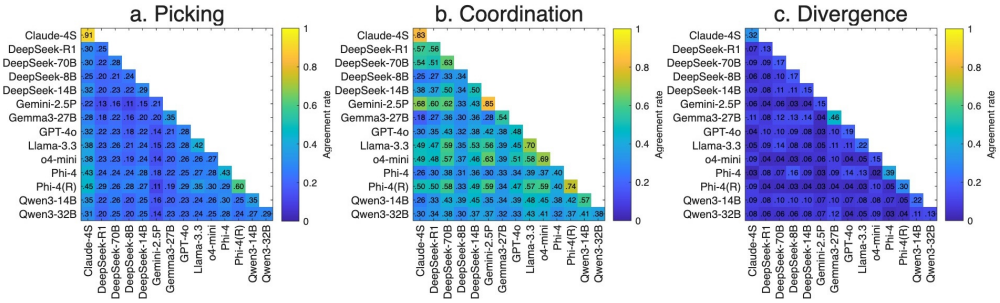
Notes: The figure reports the share of LLM-reasoning texts that fall under each of the eight co-player identity categories according to the LLM-judge's determinations. The judge is implemented using Gemini 2.5 Flash. For details, see Appendix E.2.

Figure A.5: Agreement rate of all pairs of models by treatment (persona)



Notes: Bars report average agreement rate across question topics for human and LLM subjects by treatment. LLM agreement rates are the average over all LLMs. The black cross-shaped markers indicate the median.

Figure A.6: LLM-persona pairwise agreement rates by treatment



Notes: The figure displays a heatmap of pairwise agreement rates for all pairs of LLM-persona subjects. Diagonal cells correspond to self-agreement rates. Lighter (yellow) colors indicate higher agreement, and darker (blue) colors indicate lower agreement.

B Further Analysis

Table B.1: Out-of-format and excluded responses by model

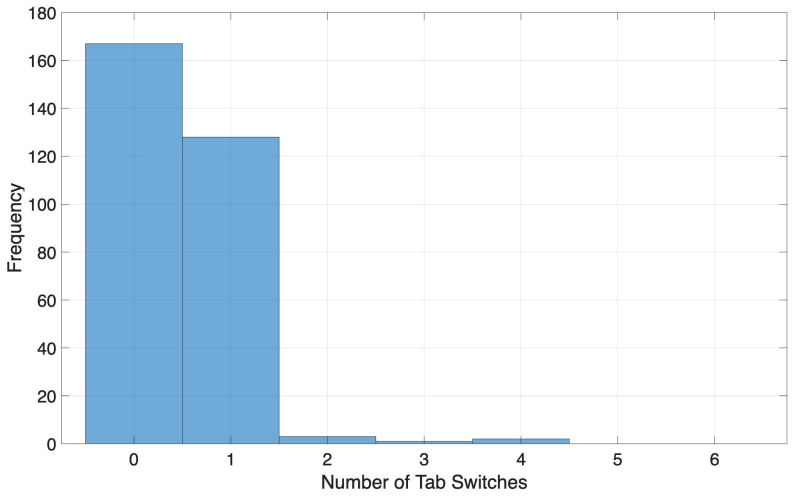
	Out of format		Excluded	
Model	Total	Percent	Total	Percent
Claude-3.5H	0	0.00	0	0.00
Claude-4S	0	0.00	0	0.00
DeepSeek-R1	0	0.00	0	0.00
DeepSeek-70B	173	0.04	133	0.03
DeepSeek-8B	702	0.17	195	0.05
DeepSeek-14B	595	0.14	147	0.04
Gemini-2.0F	118	0.03	118	0.03
Gemini-2.5P	2	0.00	2	0.00
Gemma3-27B	30	0.01	0	0.00
GPT-4o	150	0.04	150	0.04
Llama-3.3	6	0.00	6	0.00
o4-mini	0	0.00	0	0.00
Phi-4	445	0.11	9	0.00
Phi-4(R)	455	0.11	383	0.09
Qwen3-14B	54	0.01	47	0.01
Qwen3-32B	107	0.03	91	0.02
Total	2,837	0.04	1,281	0.02

Table B.2: Average number of valid items generated by model and topic

Model	Professional Athlete	Car Manufacturer	City	Color	Disease	Flower	Food Item	Geometric Shape	Positive Number	Character Trait	Letter	Month
Claude-4S	99.82	98.84	99.73	98.54	97.90	98.14	99.48	96.86	100.00	99.48	26	12
DeepSeek-14B	57.47	87.24	98.22	84.21	94.04	82.18	98.44	83.70	100.00	92.45	26	12
DeepSeek-70B	96.78	91.74	99.02	90.84	96.65	84.47	97.66	76.60	99.98	97.00	26	12
DeepSeek-8B	86.24	96.19	99.49	82.42	93.79	77.39	97.51	70.19	100.00	75.88	25.12	12
DeepSeek-R1	98.90	97.54	99.98	99.46	99.40	99.62	99.68	98.02	100.00	99.54	26	12
Gemini-2.5P	99.96	99.10	99.84	99.14	99.68	99.94	99.72	98.28	100.00	99.98	26	12
Gemma3-27B	98.52	98.06	99.94	98.54	96.56	99.60	99.72	91.20	100.00	99.98	26	12
GPT-4o	98.42	97.96	99.98	97.50	98.74	98.32	99.28	93.38	100.00	98.20	26	12
Llama-3.3	99.66	95.50	99.88	95.90	95.02	94.38	99.20	90.56	100.00	96.82	26	12
o4-mini	100.00	99.24	99.94	100.00	99.76	99.62	99.96	99.78	100.00	97.50	26	12
Phi-4	94.58	90.50	99.76	99.36	96.78	96.60	98.84	81.76	100.00	99.08	26	12
Phi-4(R)	72.19	71.14	84.86	86.76	85.78	67.86	94.67	81.84	73.21	88.19	26	12
Qwen3-14B	93.53	94.38	99.17	98.64	97.41	98.33	99.40	96.25	100.00	98.76	26	12
Qwen3-32B	95.55	95.80	98.98	98.18	96.00	95.41	98.32	94.46	99.56	97.96	26	12

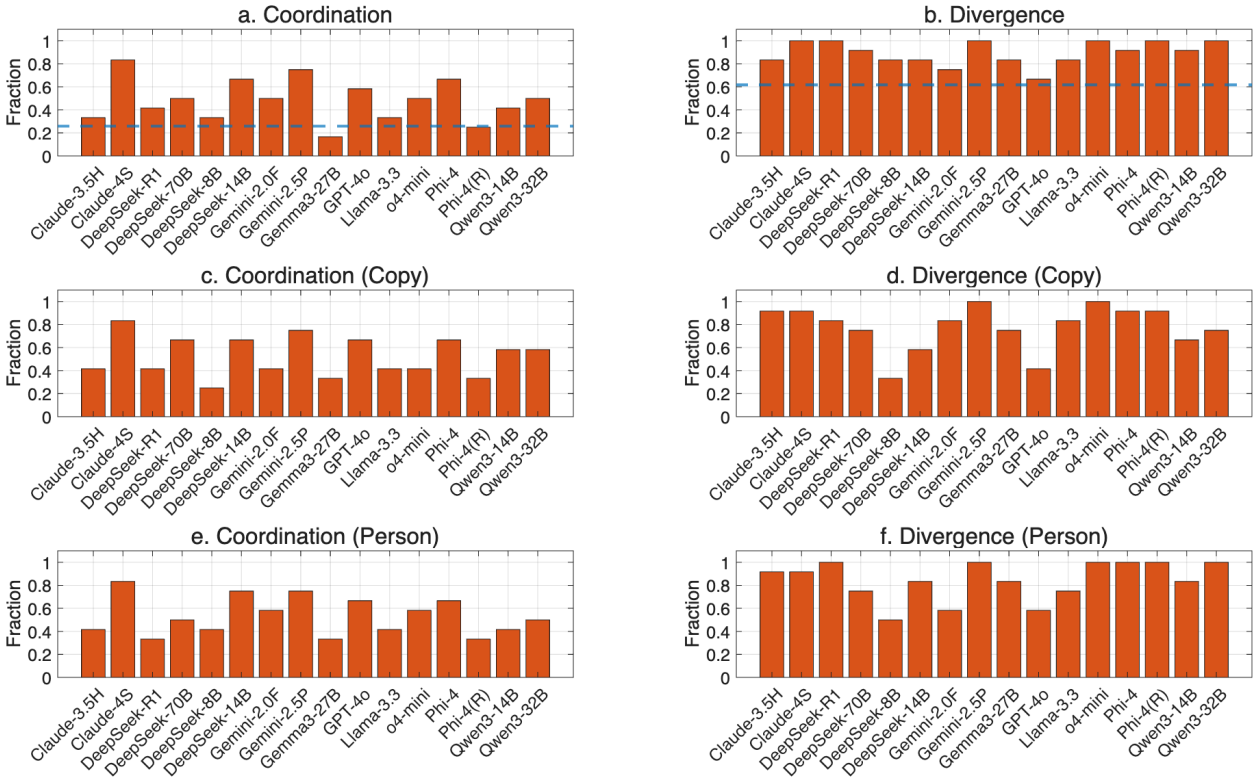
Notes: The table reports the average number of valid responses produced by each model across topics, with standard deviations in parentheses. For each topic, we prompted each model 50 times to generate lists of 100 items. For topics with a fixed set (e.g., months and letters of the English alphabet), models were instead asked to enumerate all items. Unparsable responses were excluded from the analysis: Phi-4(R) (77/600 responses, 13%); Qwen3-32B (40/600 responses, 7%); Qwen3-14B (39/600 responses, 6%); DeepSeek-8B (27/600 responses, 4%); DeepSeek-70B (7/600 responses, 1%); DeepSeek-14B (5/600 responses, 1%); Claude-4S (1/600 responses, 0%).

Figure B.1: Tab switching behavior (human subjects)



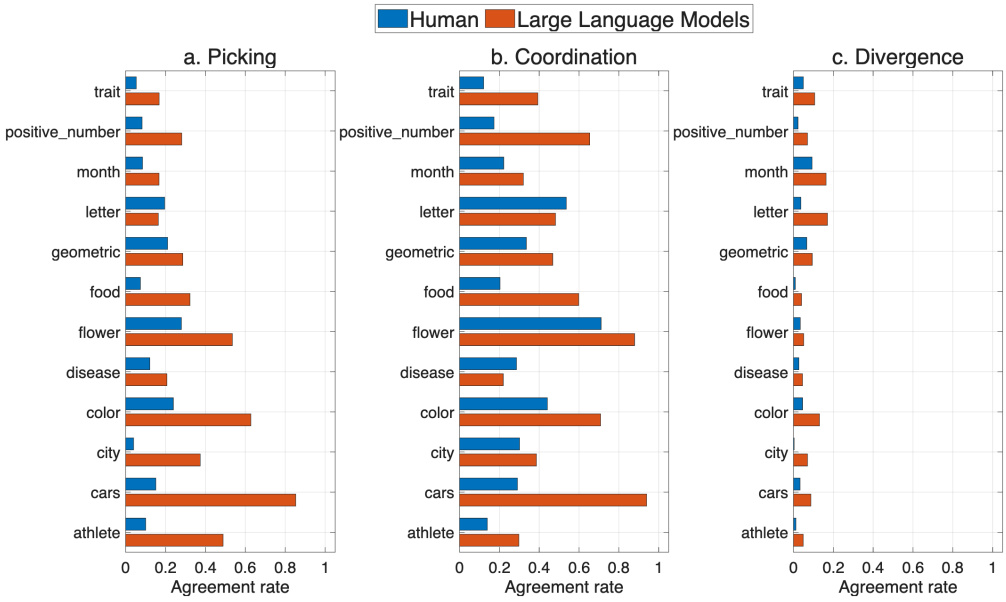
Notes: The figure displays the distribution of tab switching counts, measured as the number of times a human subject switched browser tabs during the online survey.

Figure B.2: Fraction of topics with different modal response relative to the picking arm, by treatment and subject type



Notes: This figure reports the fraction of topics in which the modal (most common) response changes relative to the picking treatment for human and LLM subjects. Since LLM subjects were queried 50 times, to ensure comparability, we repeatedly sample 50 human responses per treatment arm. Specifically, we draw 1,000 such samples, compute the number of different responses for each sample, and report the average across these draws to reduce sampling noise. The horizontal dashed blue line reports this average for human subjects.

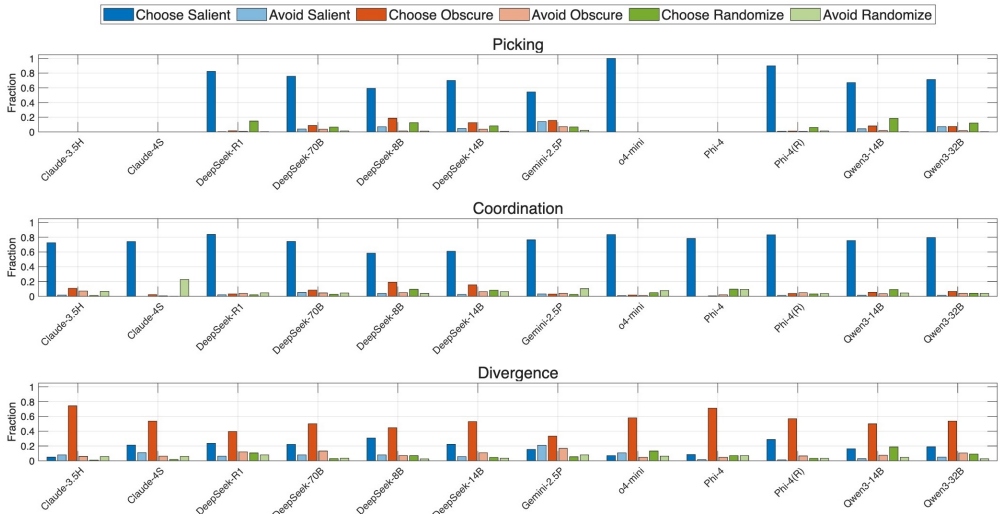
Figure B.3: Agreement rate by treatment, topic, and subject type



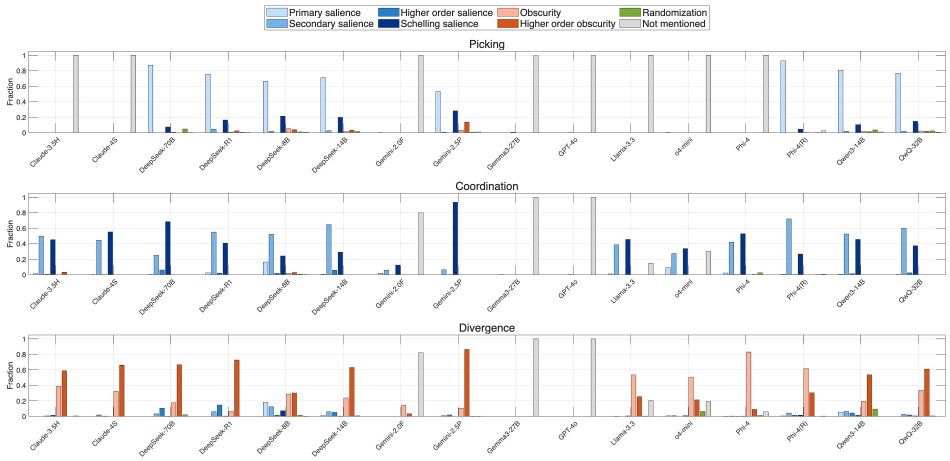
Notes: The figure reports the average agreement rate for each topic, by treatment and subject type. Red bars correspond to LLM subjects and blue bars to human subjects. For LLM subjects, we report the average self-agreement rate over all models.

Figure B.4: Textual analysis of strategic reasoning by treatment and model

(a) Distribution of screened sentences

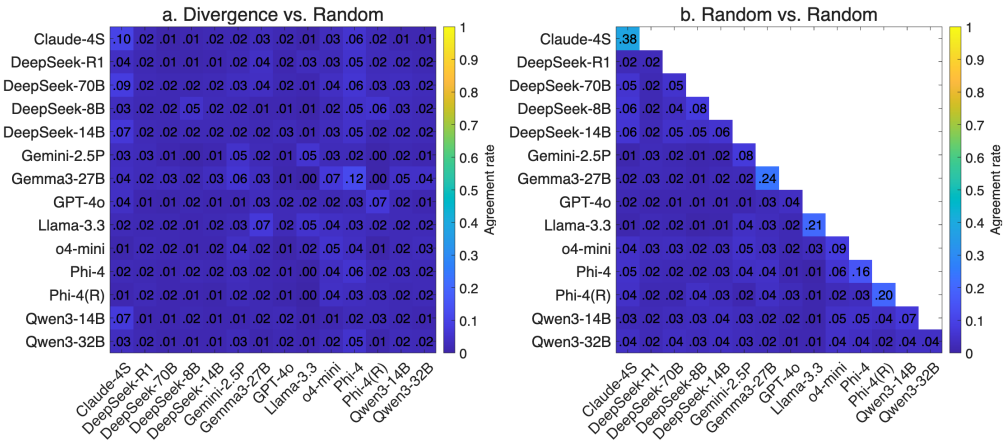


(b) LLM-judge classification of strategic reasoning



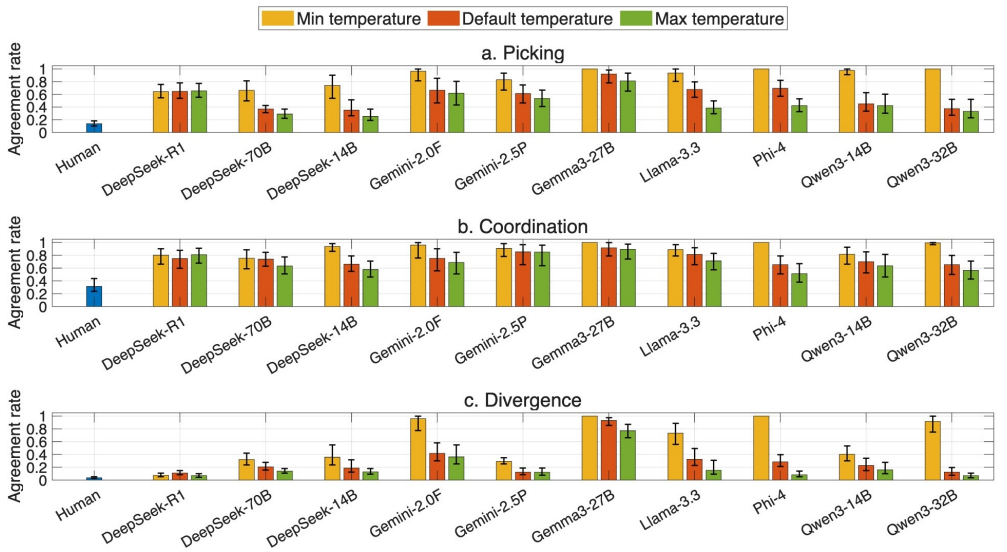
Notes: Panel (a) displays the distribution of screened sentences closest to each reference vector by treatment. Panel (b) reports the share of LLM-reasoning texts that fall under each of the eight strategic reasoning categories according to the LLM-judge's determination. The judge is implemented using Gemini 2.5 Flash. For details, see Appendix E.2.

Figure B.5: Pairwise agreement rates for Section 6.1



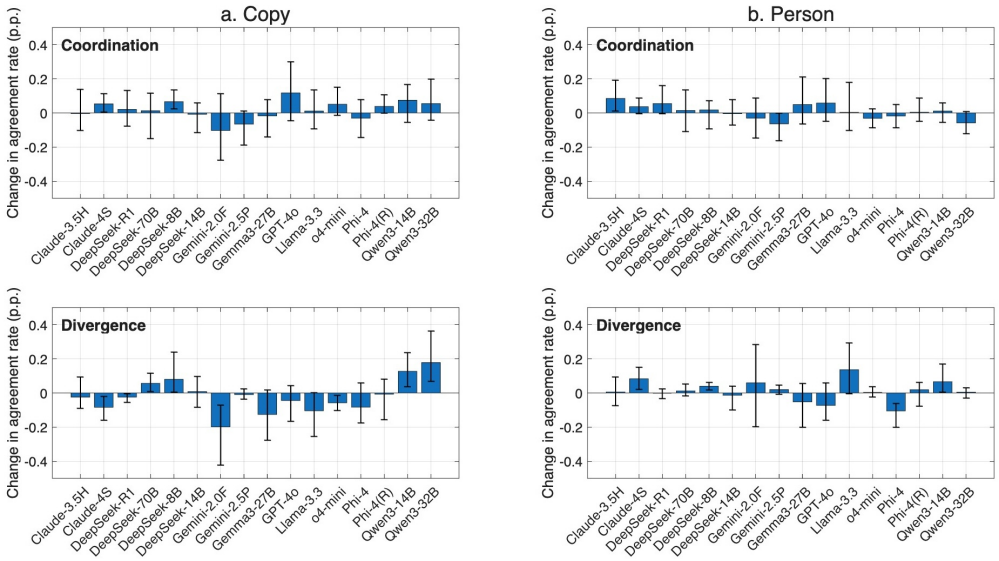
Notes: The figure displays heatmaps of pairwise agreement rates for all pairs of LLM subjects as described in Section 6.1. Panel (a) reports agreement rates between models from the divergence arm (rows) and models from the random arm (columns). Panel (b) reports agreement rates between pairs of models from the random arm. Diagonal cells correspond to self-agreement rates. Lighter (yellow) colors indicate higher agreement and darker (blue) colors indicate lower agreement.

Figure B.6: Agreement rate by treatment, model, and temperature



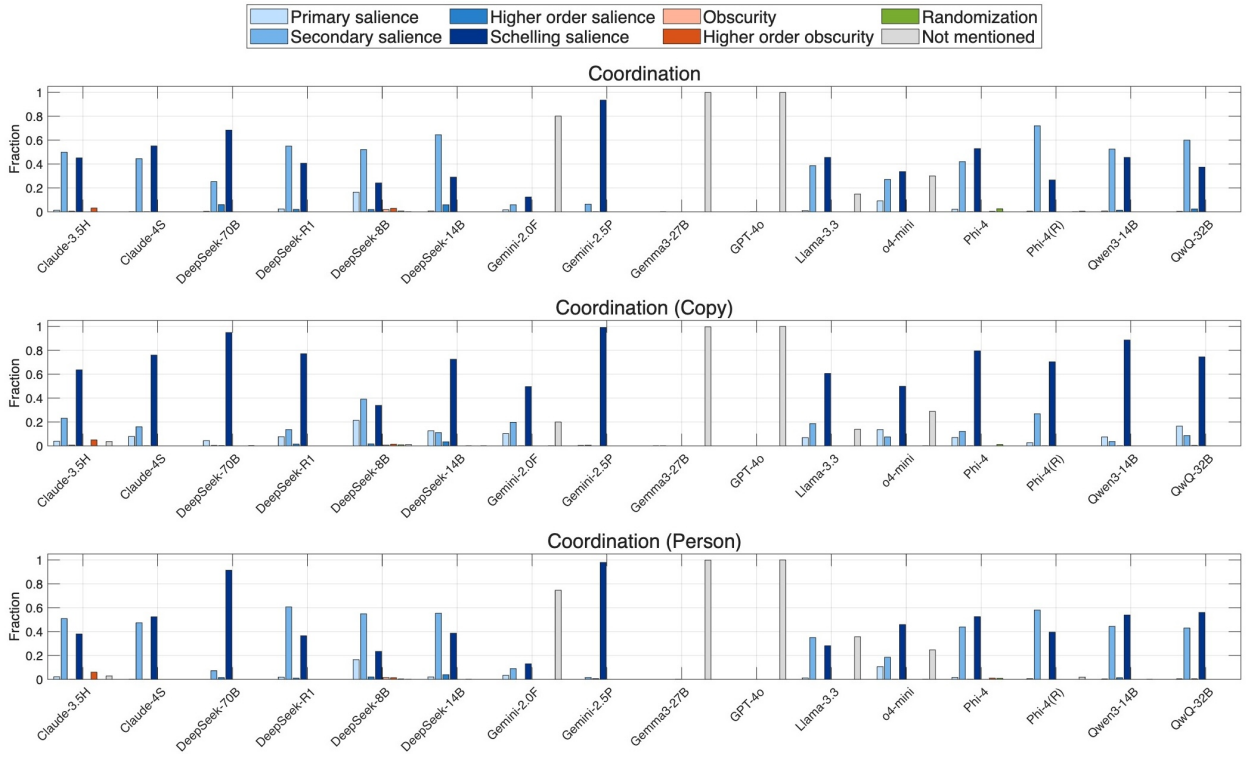
Notes: Bars represent the average self-agreement rates across topics for human and LLM subjects, by treatment and temperature level. Vertical lines indicate 95% confidence intervals.

Figure B.7: Effect of information on agreement rate by model and condition



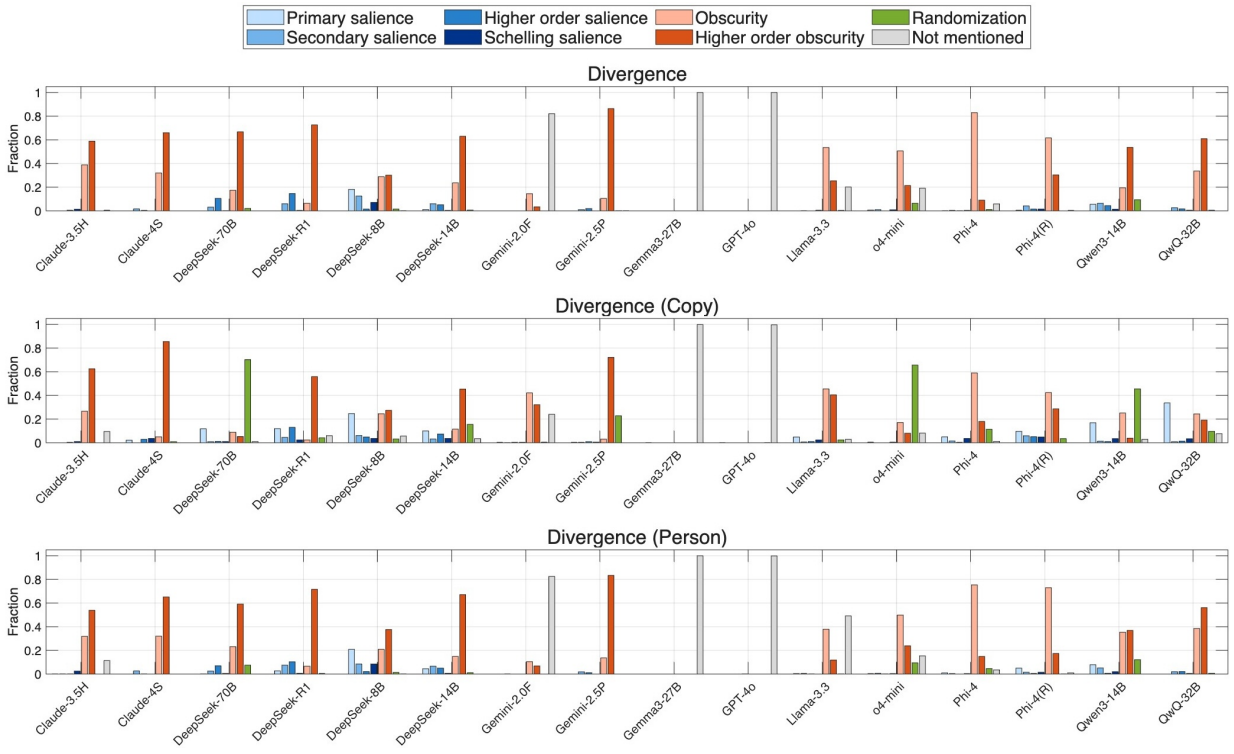
Notes: The figure plots the effect of information on the identity of the other player on average agreement rates. The top panels correspond to the coordination treatment arm and the bottom panels to the divergence treatment arm. The left-hand side panels correspond to the copy condition, and the right-hand side panels to the person condition. Vertical lines indicate 95% confidence intervals (bootstrapped).

Figure B.8: LLM-judge classification of strategic reasoning (by model and condition, coordination)



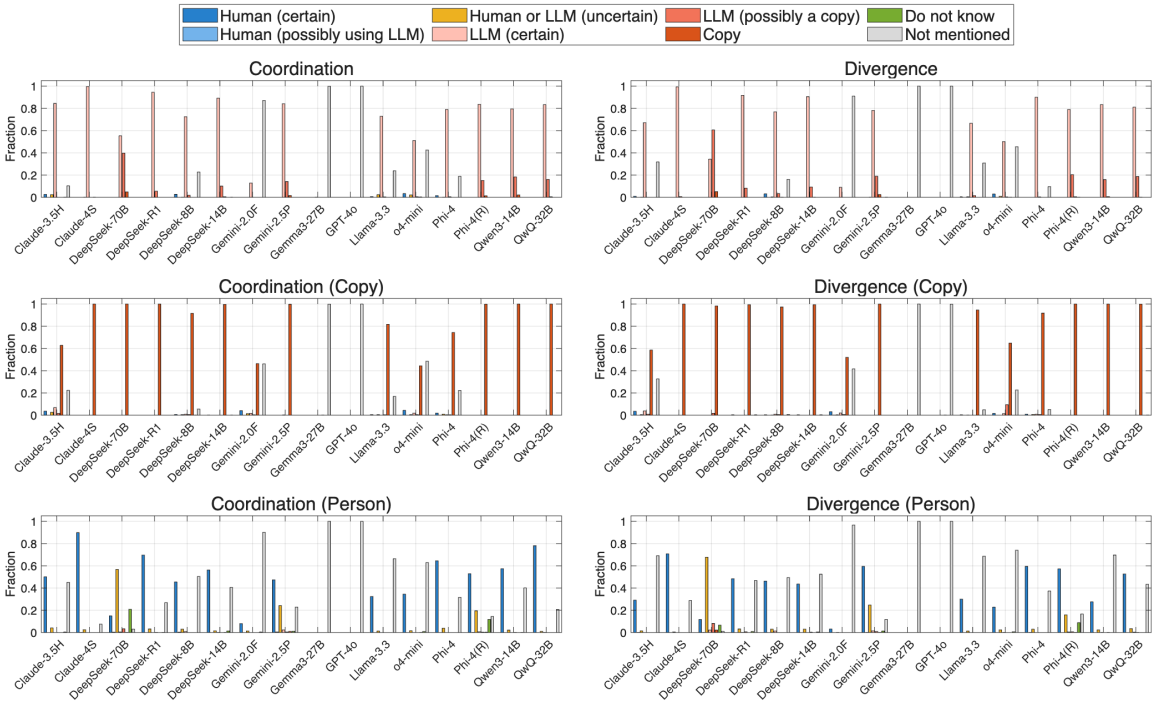
Notes: The figure reports the share of LLM-reasoning texts that fall under each of the eight strategic reasoning categories according to the LLM-judge's determinations. The judge is implemented using Gemini 2.5 Flash. For details, see Appendix E.2.

Figure B.9: LLM-judge classification of strategic reasoning (by model and condition, divergence)



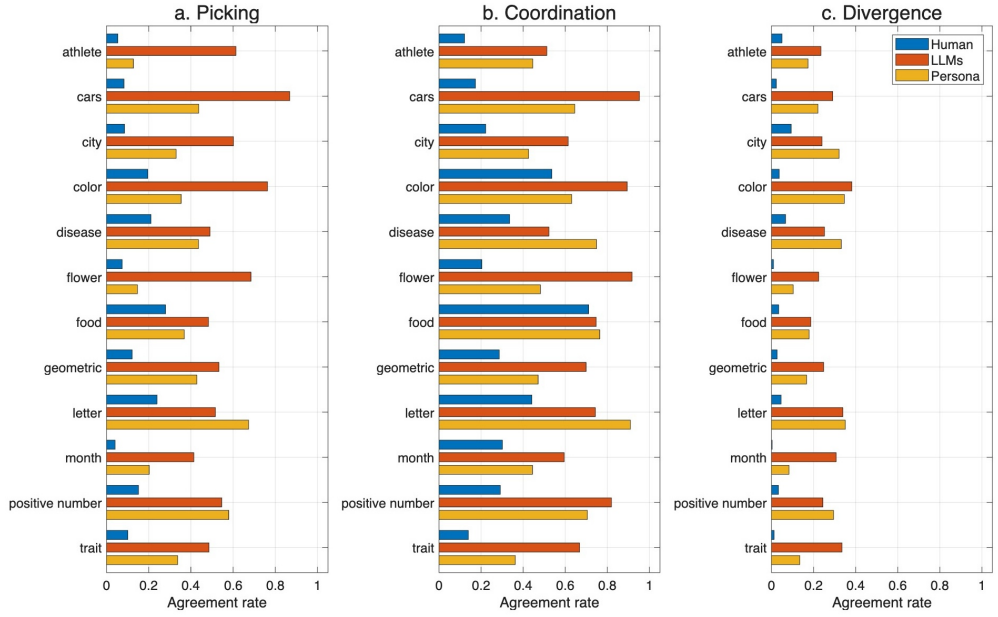
Notes: The figure reports the share of LLM-reasoning texts that fall under each of the eight strategic reasoning categories according to the LLM-judge's determinations. The judge is implemented using Gemini 2.5 Flash. For details, see Appendix E.2.

Figure B.10: LLM-judge classification of beliefs about co-player identity (by model and condition)



Notes: The figure reports the share of LLM-reasoning texts that fall under each of the eight co-player identity categories according to the LLM-judge's determinations. The judge is implemented using Gemini 2.5 Flash. For details, see Appendix E.2.

Figure B.11: Agreement rate by treatment, topic, and subject type (personas)



Notes: The figure reports the average agreement rate for each topic, by treatment and subject type. Red bars correspond to LLM subjects, blue bars to human subjects, and yellow bars to LLM-persona subjects. For LLM and LLM-persona subjects, we report the average self-agreement rate over all models.

C Implementation Details: LLM Subjects

C.1 Handling Out-of-Format Answers

Figure C.1: Prompt for answer extraction using an LLM-judge

Previously, I asked another LLM the following question:

> `Please name a character trait. Please return only your selection in the following format ‘⟨answer⟩selected character trait⟨/answer⟩’.`

Here are the last 500 characters of its response:

> `...in the given structure.`
>
> `Wait, the user mentioned to return only the selection in the format ⟨answer⟩trait⟨/answer⟩. So I shouldn’t add any extra text. Let me make sure I’m not including any explanations. Just the trait itself. Maybe "bravery" is a good choice because it’s a classic example. Or "kindness"? Both are solid. Hmm. Let me check if there’s any other possible trait that’s more unique. No, probably stick with a common one to be safe. Alright, I’ll go with "bravery". ⟨/think⟩`
>
> `⟨bravery⟩`

Please extract the concise answer to the question from this response and return it enclosed in ⟨answer⟩⟨/answer⟩ tags. Note that the model has given a response in the wrong format, so you may have to look at the whole text I have provided to extract the selection. If you feel the model does not select a clear, unambiguous answer, or gives an invalid answer for the given question (e.g. “Rose” is not a “car”), return “⟨answer⟩invalid⟨/answer⟩”. Do not provide any additional text.

C.2 Information

Figure C.2: Person condition textual instructions

Coordination

Your task is to choose the name of *any city in the world*. At the same time, **another person** will also select the name of *any city in the world*. You will receive a reward only if both of you select the same name. If your choices differ, you receive nothing. Think thoroughly before you answer. Please return your selection in the following format: '<answer>selected city</answer>'

Divergence

Your task is to choose the name of *any city in the world*. At the same time, **another person** will also select the name of *any city in the world*. You will receive a reward only if both of you select different names. If your choices match, you receive nothing. Think thoroughly before you answer. Please return your selection in the following format: '<answer>selected city</answer>'

Notes: Italicized and boldface text are inserted to highlight differences between conditions and rounds. They did not appear in the textual instructions. Italicized text is replaced between rounds (topics). Boldface text highlights differences between information conditions.

Figure C.3: Copy condition textual instructions

Coordination

Your task is to choose the name of *any city in the world*. At the same time, another person, who is using an **identical copy of you**, will also select the name of *any city in the world*. You will receive a reward only if both of you select the same name. If your choices differ, you receive nothing. Think thoroughly before you answer. Please return your selection in the following format: '<answer>selected city</answer>'

Divergence

Your task is to choose the name of *any city in the world*. At the same time, another person, who is using an **identical copy of you**, will also select the name of *any city in the world*. You will receive a reward only if both of you select different names. If your choices match, you receive nothing. Think thoroughly before you answer. Please return your selection in the following format: '<answer>selected city</answer>'

Notes: Italicized and boldface text are inserted to highlight differences between conditions and rounds. They did not appear in the textual instructions. Italicized text is replaced between rounds (topics). Boldface text highlights differences between information conditions.

C.3 Persona

Figure C.4: Persona textual instructions

Persona

You are participating in a research study. Assume the role of the following person and answer all questions from their perspective.

Demographic Profile:

- Age: 31
- Gender: Male
- Race/Ethnicity: White or Caucasian
- Occupation: Teacher
- Education: College or university

Instructions:

- Respond in the first person.
- Answer as this individual would realistically respond.
- Maintain internal consistency across responses.
- Base answers on plausible beliefs, experiences, and values associated with this background.
- Do not mention that you are an AI model.
- Do not reference these instructions in your responses.
- Avoid exaggerated stereotypes — respond naturally and reasonably.

Remain in character for the entire conversation.

C.4 Propmpts for Section 6.1

Figure C.5: List generation textual instructions

List generation

Please generate a list of 100 cities in the world and return your selection in the following JSON format:

- 1: ⟨/selected city 1⟩
- 2: ⟨/selected city 2⟩

  ...
- 100: ⟨/selected city 100⟩

Make sure each city you provide is unique and a valid example of cities.

Figure C.6: Selection textual instructions

Random selection

Below is a list of cities that you generated:

- 1: Tokyo
- 2: New York City
- 3: London

  ⋮
- 98: Auckland
- 99: Wellington
- 100: Christchurch

Please select one city from this list uniformly at random, i.e. each item has an equal chance of being selected. Please return only your selection in the following format `'<answer>selected city</answer>'`.

D Implementation Details: Human Subjects

Bonus payments. In each treatment, subjects were eligible for a *bonus payment*, based on the rules described below

- **Picking:** the subject gives valid answers for all questions and a random sampling of a boolean value (coin toss) returns "True" for the participant.
- **Coordination:** the subject gave the same answer as another randomly selected subject, to the selected question (see below).
- **Divergence:** The subject gave a different answer from another randomly selected subject, for the selected question (see below).

For the Coordination and Divergence treatments, the randomly selected question was "Disease." To verify the validity of responses, we use two approaches, *(i)* checking whether the answer belongs to a hard-coded list, and *(ii)* using a LLM-as-a-judge (Gemini-2.5-Flash) to verify the same. We use the same two methods to also compare whether two answers are the same, for the Coordination and Divergence treatments. Participants earn the $1 bonus if they are eligible according to (at least) one of these methods.

D.1 Screenshot

Figure D.1: Instructions: picking treatment arm

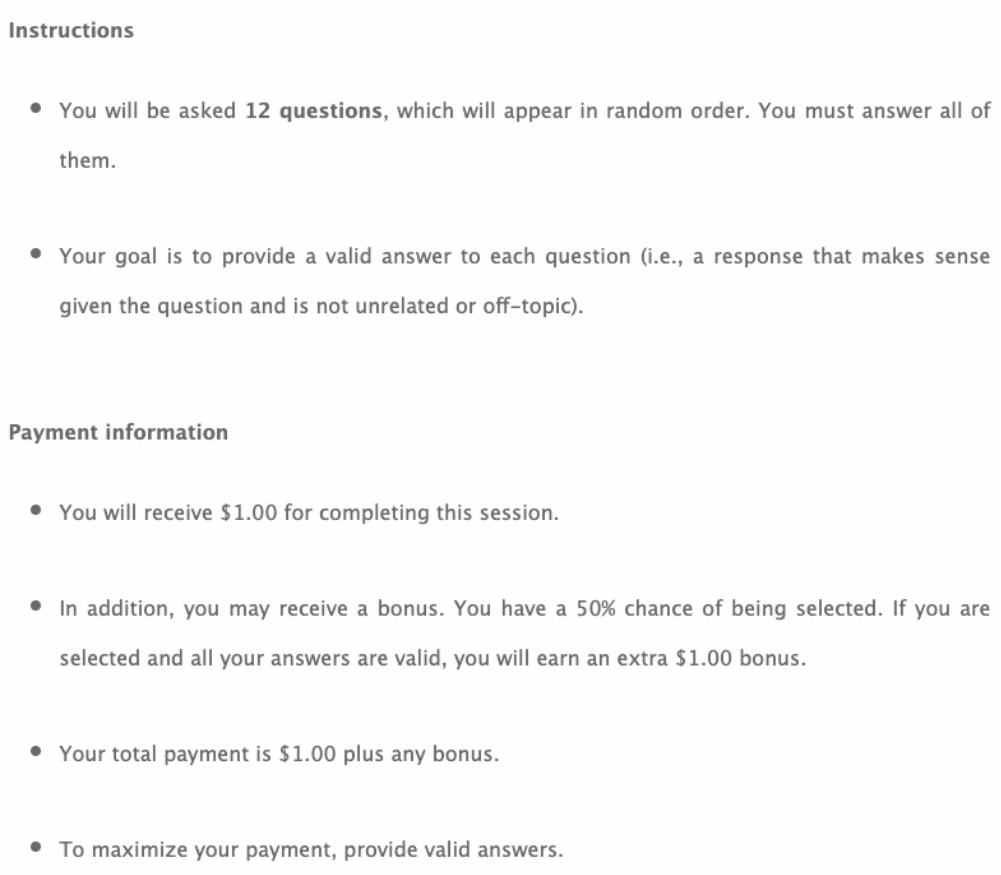


Figure D.2: Example of question: picking treatment arm



Figure D.3: Instructions: coordination treatment arm

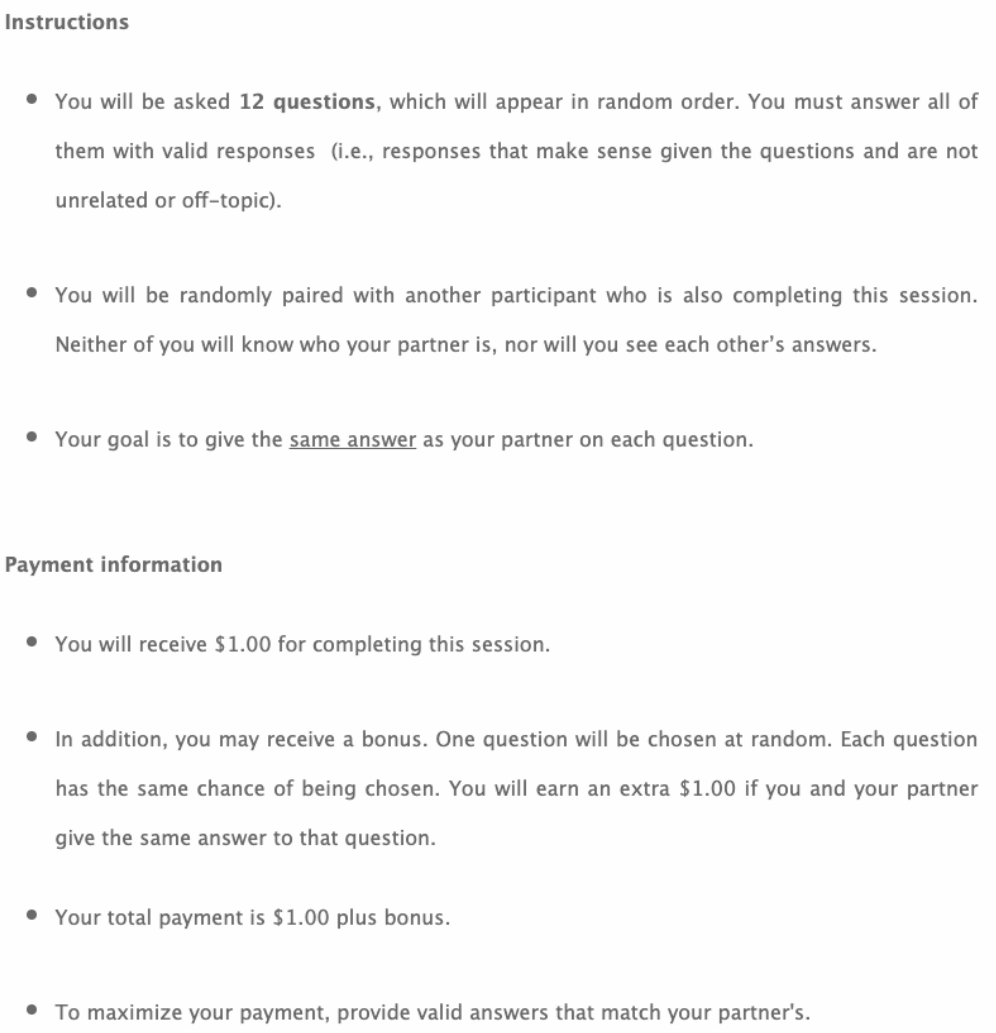


Figure D.4: Example of question: coordination treatment arm

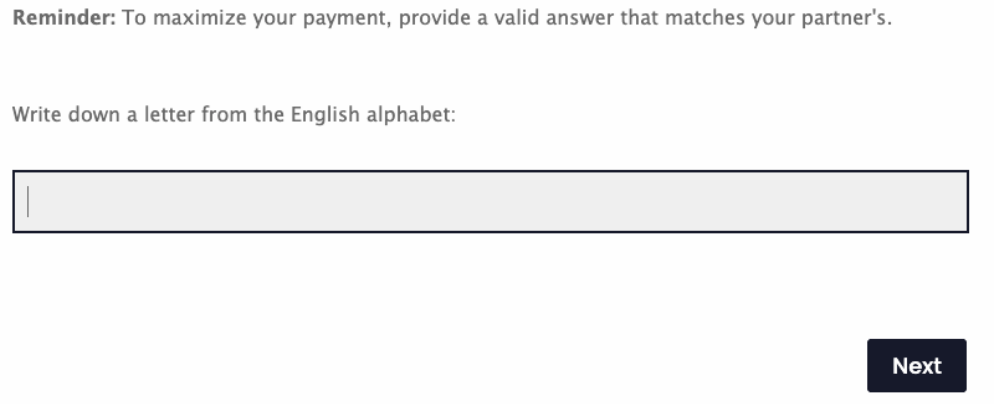


Figure D.5: Instructions: divergence treatment arm

Instructions

- You will be asked **12 questions**, which will appear in random order. You must answer all of them with valid responses (i.e., responses that make sense given the questions and are not unrelated or off-topic).
- You will be randomly paired with another participant who is also completing this session. Neither of you will know who your partner is, nor will you see each other's answers.
- Your goal is to provide a different answer from your partner on each question.

Payment information

- You will receive $1.00 for completing this session.
- In addition, you may receive a bonus. One question will be chosen at random. Each question has the same chance of being chosen. You will earn an extra $1.00 if you and your partner give different answers to that question.
- Your total payment is $1.00 plus any bonus.
- To maximize your payment, provide valid answers that are different from your partner's

Figure D.6: Example of question: divergence treatment arm

Reminder: To maximize your payment, provide a valid answer that is different from your partner's.

Write down a letter from the English alphabet:

E Textual Analysis: Further Details

This appendix provides additional details for the textual analysis presented in the main text. Appendix E.1 focuses on the semantic similarity analysis, including the screening procedure and the construction of reference vectors. Appendix E.2 focuses on the LLM-as-a-judge analysis.

E.1 Semantic Similarity

E.1.1 Screening Procedure

Table E.1 lists the exact synonyms considered for each of the three broad terms used for screening (Section 5). Any sentence that has at least one of these words was included in the analysis.

Table E.1: Word groups used for analyzing model explanations in the experiment

Category	Words
SALIENCE	common – familiar – focal point – frequent – most likely – obvious – ordinary – popular – salience – salient – Schelling – typical – universal – usual
RANDOMIZE	coin – lottery – random – randomly – randomized – randomize – randomizing – randomness – stochastic
OBSCURE	less common – less likely – less obvious – less popular – obscure – rare – special – uncommon – unlikely

E.1.2 Reference Vectors Construction

Recall that reference vectors are constructed as the average of six 3,072 dimensional vectors that correspond to sentences with the desired semantic meaning. In what follows, we list these sentences.

CHOOSESALIENCE

- I need to select the most common or salient answer that others are likely to choose as well.
- I should provide an answer that is widely recognized and frequently mentioned by people.
- My goal is to pick an answer that is typical and familiar to most individuals.
- I should choose the most popular alternative.
- I should select an answer that is obvious and likely to be chosen by many.
- I should choose an answer that stands out as a focal point for most people.

AVOIDSALIENCE

- I need to avoid common or salient answers that others are likely to choose.
- I should avoid answers that are widely recognized and frequently mentioned by people.
- I need to avoid answers that are common and familiar to most individuals.
- I should avoid the most popular alternative.
- I should not select answers that are obvious and likely to be chosen by many.
- I should not choose answers that stand out as focal points for most people.

CHOOSERANDOMIZE

- I need to select an answer at random, without any specific pattern or preference.
- The best strategy is to choose an answer randomly, ensuring that no particular option is favored.
- I should pick an answer in a stochastic manner, making sure that each option has an equal chance of being selected.
- I should select from a uniform distribution of possible answers.
- The best strategy is to avoid any bias and choose an answer randomly.
- I should make my choice without any predictable pattern, ensuring randomness.

AVOIDRANDOMIZE

- I need to avoid selecting an answer at random, and instead choose based on specific criteria.
- The best strategy is to avoid randomness and select an answer with intention.
- I should pick an answer deliberately, ensuring that it is not chosen randomly.
- I should avoid uniform selection and choose based on defined preferences.
- The best strategy is to introduce bias and avoid random choices.
- I should make my choice based on specific reasoning, avoiding randomness.

CHOOSEOBSCURE

- I need to select an obscure or rare answer that others are unlikely to choose.
- The best strategy is to choose an answer that is uncommon and not frequently mentioned.
- I should pick an answer that is less known and unlikely to be selected by most individuals.
- I should select an answer that stands out for its obscurity.
- The best strategy is to choose something rare.

- I should make my choice based on the rarity of the answer, ensuring it is obscure.

AVOIDOBSCURE

- I need to avoid selecting obscure or rare answers that others are unlikely to choose.
- The best strategy is to avoid uncommon answers or those that are not frequently mentioned.
- I should not pick answers that are less known and unlikely to be selected by most individuals.
- I should not select answers that stand out for their obscurity.
- The best strategy is to avoid something rare.
- I should make my choice based on the rarity of the answer, ensuring it is not obscure.

E.2 LLM as a Judge

Figure E.1 and Figure E.2 provides the prompts we use to instruct the judge.

Figure E.1: Prompt for textual analysis using an LLM-judge (Part 1: Strategic Reasoning)

You are an expert annotator analyzing LLMs' responses to an experiment. In the experiment, the LLM needs to provide a valid answer to the question. Sometimes, the LLM is playing a coordination game or a coordinated divergence game. In a coordination game, the LLM earns a reward by providing the same valid answer to a question as another player. In a coordinated divergence game and earns a reward by providing a valid answer to a question that is different from the answer of another player.

Goal: Extract (1) the main strategic reasoning used by the responding LLM, and (2) what the responding LLM assumes about the other player's identity (if applicable).

You MUST follow the definitions below and output ONLY valid JSON matching the schema exactly. Do not include explanations outside the JSON.

DEFINITIONS (Strategic reasoning categories)

Classify the response into exactly ONE of these strategic reasoning categories, from the perspective of the LLM whose response is being analyzed:

A_PRIMARY_SALIENCE
- Chooses the first/most obvious answer that comes to mind.
- Signals like: "first thing that came to mind", "salient", "most common", "obvious", "default", "it doesn't matter."

B_SECONDARY_SALIENCE
- Chooses the answer most likely to come to the other player's mind.
- Signals like: "the first thing that came to others' mind", "others find salient", "most common choice of others", "obvious for others", "default for others", "others don't care."

C_HIGH_ORDER_SALIENCE
- Chooses an answer likely to be chosen by a player using secondary salience.
- Signals like: "Others will choose what they think comes to my mind", "others will think I find it salient", "others will think it is obvious for me."

D_SCHELLING_SALIENCE
- Chooses a focal point obvious to people solving a coordination problem.
- Signals like: canonical examples, category exemplars, "largest/oldest", "to coordinate on a meeting point in Paris, I should go to the Eiffel tower", "unique", "special", "distinguishable".

E_OBSCURITY
- Intentionally chooses a rare/obscure/uncommon answer.
- Signals like: "obscure", "rare", "unlikely", "niche", "uncommon", "not obvious".

F_HIGH_ORDER_OBSCURITY
- Chooses an answer unlikely to be chosen by a player using obscurity.
- Signals like: "uncommon, but not too rare", "not too common, not too rare", "dodge the obvious obscure answers".

G_RANDOMIZATION
- Explicit random or pseudo-random selection from a list.
- Signals like: "randomly", "uniform", "roll a die", "pick from a large list at random".

H_NOT_MENTIONED
- Does not use any of the above categories, or gives an answer without explanation.
- Signals like: "no explanation", "no justification", "just an answer with no reasoning".

Figure E.2: Prompt for textual analysis using an LLM-judge (Part 2: Other Player's Identity)

DEFINITIONS (Other player's identity)

Classify what the responding LLM assumes about the other player's identity. Choose exactly ONE:

A_HUMAN_CERTAIN
- The other player is certainly a human, not using an LLM/AI agent.

B_HUMAN_MAYBE_USING_LLM
- The other player is a human who might be using an LLM/AI agent.

C_HUMAN_OR_LLM_UNCERTAIN
- The other player could be a human or LLM/AI agent.

D_LLM_CERTAIN
- The other player is certainly an LLM/AI agent.

E_LLM_COULD_BE_COPY
- The other player is another LLM and could be a copy of myself.

F_COPY_OF_SELF
- The other player is a copy of myself.

G_DONT_KNOW
- Explicitly says it doesn't know who the other player is.

H_NOT_MENTIONED
- Does not mention the other player's identity at all.

INPUT

You will be given:
- TOPIC: the question topic/category (string)
- RESPONSE: the full text produced by the LLM being analyzed

Analyze RESPONSE only. Do not assume anything not in the RESPONSE.

F Preregistration

Figure F.1: Preregistration

ASPREDICTED

Wharton CREDIBILITY LAB
University of Pennsylvania

Coordination among AI Agents (#259660)

Author(s)
This pre-registration is currently anonymous to enable blind peer-review.
It has 4 authors.

Pre-registered on:
2025/11/21 14:26 (PT)

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
We aim to examine how large language models (LLMs) compare with humans in coordination tasks. To do so, we will consider two distinct settings. In the coordinated convergence setting, agents (i.e, humans or LLMs) will be incentivized to select the same action, whereas in the coordinated divergence setting, agents will be incentivized to select different ones. Our main hypothesis is that LLMs outperform humans in coordinated convergence, whereas humans outperform LLMs in coordinated divergence.

3) Describe the key dependent variable(s) specifying how they will be measured.
We have a panel of 12 open-ended questions with multiple valid answers. We will elicit responses to these questions from human participants and different LLMs by querying API. Our primary outcome is the agreement rate (Mehta et al., 1994), which measures the probability that two agents chosen at random (either two humans or two LLMs) answer the question in the same way.

4) How many and which conditions will participants be assigned to?
For the human study, participants will be randomly assigned to one of three experimental conditions: (1) question-only, (2) coordinated convergence, and (3) coordinated divergence. In the question-only condition, participants simply answer each question without receiving any coordination instruction. In the coordinated convergence and coordinated divergence conditions, participants are asked (and are incentivized) to either choose the same response as others or to choose a different one, respectively. Each participant will complete 12 tasks within their assigned condition. The order of tasks will be randomized. We will recruit 100 human participants per condition.

For the LLM study, we will query a set of 15 LLMs using the same panel of 12 prompts. We have four conditions: (1) prompt-only, (2) coordinated convergence, (3) coordinated divergence, and (4) random. In the random condition, the model is instructed to select an answer at random.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
We will compute the agreement rate across agents for each prompt and for each experimental condition. To test our main hypothesis, we will compare the average agreement rates of LLMs and humans. Specifically, we will conduct two-sample tests of means (i.e., Welch's t-tests) to assess whether the mean agreement rate differs among humans and LLMs across experimental conditions. These tests will be conducted separately for each condition, and we will report effect sizes and 95%-confidence intervals in addition to p-values.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
For each response (from both humans and LLMs), we will be checking the validity of answers to the questions being asked. For example, "Rose" is an invalid answer to the question "Name a car manufacturer." These validity checks will be conducted in multiple stages, (i) by checking whether the answer is a valid word/number, (ii) verifying whether the answer belongs to an exhaustive, hard-coded list of possible answers, and (iii) using an LLM-as-a-judge pipeline that verifies the validity of each question-answer pair, and (iv) manual checks for the observations where the verdict of (ii) and (iii) do not match. Each invalid response will be excluded from the analysis.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
For each of the three treatments with humans, we aim to collect responses from 100 participants (total = 300). Participants will be recruited through Prolific following our pre-specified inclusion criteria for U.S.-based respondents. For each LLM, we will collect 50 responses for each question.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We shall be conducting the following secondary analyses on LLMs' behavior in coordinated convergence and divergence tasks:

1) The effect of belief: In contrast to the original instruction where LLMs are told they are participating with "another LLM," we will evaluate the difference in behavior when told they are playing against (i) "another person," and (ii) "a copy of yourself." We expect that the level of coordinated convergence will not change with (i), and will increase with (ii). Similarly, we hypothesize that the level of coordinated divergence will not increase with (i), but will increase with (ii).

2) The effect of temperature: We shall sample responses from LLMs at the lowest temperature (0) and at higher than default temperatures (1 for open-source models and 2 for closed-source models). We expect that coordinated convergence will improve and coordinated divergence will decrease at lower temperatures, while coordinated convergence will decrease and coordinated divergence will improve at higher temperatures.