Supplementary Information for

Assessing mentalization in humans and large language models

Aamir Sohail[a,b,c*], Xintong Zhong[d], Arkady Konovalov[a,b], Patricia L. Lockwood[a,b,c,e,f], Lei Zhang[a,b,c,e*]

[a]Centre for Human Brain Health, University of Birmingham, Birmingham, UK.

[b]School of Psychology, University of Birmingham, Birmingham, UK.

[c]Institute for Mental Health, University of Birmingham, Birmingham, UK.

[d]Department of Psychology, Virginia Tech, Blacksburg, USA.

[e]Centre for Developmental Science, University of Birmingham, Birmingham, UK.

[f]Department of Experimental Psychology, University of Oxford, Oxford, UK.

*Corresponding author(s).

E-mail(s): axs2210@student.bham.ac.uk; l.zhang.13@bham.ac.uk.

Table of contents

Prompts **3**
Inspection game 3
Rock-paper-scissors 4

Computational modeling **6**
Inspection game 6
Rock-paper-scissors 7

Supplementary analyses **10**
Human/LLM analyses 10
Checks 14

Appendix **24**
Inspection game 24
Rock-paper-scissors 26

References **28**

Prompts

Inspection game

In the standard prompt condition, the following text was provided to each LLM:

> Round <CURRENT ROUND> of 150
>
> You are participating in a game where the goal of both players is to maximize their respective earnings.
>
> In the game, there are two roles, the EMPLOYER and the EMPLOYEE. Your role in this game is the EMPLOYEE.
>
> You will play 150 repeated rounds with the same EMPLOYER, whose decisions will be actively made by a HUMAN participant in real time, and NOT by a computer or artificial player.
>
> In each round the EMPLOYER will have to make a choice, whether to INSPECT or NOT INSPECT. In each round the EMPLOYEE will have to make a choice, whether to WORK or SHIRK.
>
> The rules of the game are as follows:
>
> If the EMPLOYER chooses NOT INSPECT and the EMPLOYEE chooses WORK, the EMPLOYER earns 100 cents, and the EMPLOYEE earns 0 cents.
>
> If the EMPLOYER chooses INSPECT and the EMPLOYEE chooses WORK, the EMPLOYER earns 0 cents, and the EMPLOYEE earns 50 cents.
>
> If the EMPLOYER chooses NOT INSPECT and the EMPLOYEE chooses SHIRK, the EMPLOYER earns 0 cents and the EMPLOYEE earns 50 cents.
>
> If the EMPLOYER chooses INSPECT and the EMPLOYEE chooses SHIRK, the EMPLOYER earns 25 cents, and the EMPLOYEE earns 0 cents.
>
> Below is the game history for the previous <HISTORY> rounds:
>
> In round <HISTORY ROUNDS>, the EMPLOYEE chose <EMPLOYEE ACTION> and the EMPLOYER chose <EMPLOYER ACTION>. Thus, you earned <EMPLOYEE PAYOFF> cents and the employer earned <EMPLOYER PAYOFF> cents.
>
> In total from rounds <HISTORY>:
>
> The EMPLOYER chose to INSPECT <HISTORY INSPECT> times and chose to NOT INSPECT <HISTORY NOT INSPECT> times.
>
> The EMPLOYEE chose to WORK <HISTORY WORK> times and chose to SHIRK <HISTORY SHIRK> times.
>
> You are currently playing round <CURRENT ROUND>.
>
> Given your role as the EMPLOYEE, what do you choose?
>
> WORK
>
> SHIRK

The SCoT condition provided the following text, consisting the baseline text plus an additional question to record the model's prediction of the opponent's choice. Additions to the original prompt are highlighted in bold text:

> Round <CURRENT ROUND> of 150

You are participating in a game where the goal of both players is to maximize their respective earnings.

In the game, there are two roles, the EMPLOYER and the EMPLOYEE. Your role in this game is the EMPLOYEE.

You will play 150 repeated rounds with the same EMPLOYER, whose decisions will be actively made by a HUMAN participant in real time, and NOT by a computer or artificial player.

In each round the EMPLOYER will have to make a choice, whether to INSPECT or NOT INSPECT. In each round the EMPLOYEE will have to make a choice, whether to WORK or SHIRK.

The rules of the game are as follows:

If the EMPLOYER chooses NOT INSPECT and the EMPLOYEE chooses WORK, the EMPLOYER earns 100 cents, and the EMPLOYEE earns 0 cents.

If the EMPLOYER chooses INSPECT and the EMPLOYEE chooses WORK, the EMPLOYER earns 0 cents, and the EMPLOYEE earns 50 cents.

If the EMPLOYER chooses NOT INSPECT and the EMPLOYEE chooses SHIRK, the EMPLOYER earns 0 cents and the EMPLOYEE earns 50 cents.

If the EMPLOYER chooses INSPECT and the EMPLOYEE chooses SHIRK, the EMPLOYER earns 25 cents, and the EMPLOYEE earns 0 cents.

Below is the game history for the previous <HISTORY> rounds:

In round <HISTORY ROUNDS>, the EMPLOYEE chose <EMPLOYEE ACTION> and the EMPLOYER chose <EMPLOYER ACTION>. Thus, you earned <EMPLOYEE PAYOFF> cents and the employer earned <EMPLOYER PAYOFF> cents.

In total from rounds <HISTORY>:

The EMPLOYER chose to INSPECT <HISTORY INSPECT> times and chose to NOT INSPECT <HISTORY NOT INSPECT> times.

The EMPLOYEE chose to WORK <HISTORY WORK> times and chose to SHIRK <HISTORY SHIRK> times.

Your task is to firstly predict the next move of the EMPLOYER in each round.

You are currently playing round <CURRENT ROUND>.

Given your role as the EMPLOYEE, what do you predict your HUMAN opponent, the EMPLOYER will choose, INSPECT or NOT INSPECT?

INSPECT

NOT INSPECT

Rock-paper-scissors

In the standard prompt condition, the following text was provided:

You are playing a repeated game of {C.NUM_ROUNDS} rounds against the same opponent, whose decisions

will be actively made by a HUMAN participant in real time, and NOT by a computer or artificial player.

In each round, both players will make a single choice corresponding to the rules of the game described below.

Your aim is to maximise the amount of points scored which will be converted into a bonus payment at the end of the game.

The rules of the game are as follows:

Each player must select a single letter: J, Q or Z where:

J beats Z and loses to Q

Q beats J and loses to Z

Z beats Q and loses to J

In each round, the winner gets +1 point, and the loser gets -1 point.

In the event of a tie, both players earn 0 points.

You MUST choose either J, Q, or Z.

Below is the game history for the previous <HISTORY> rounds:

In round <HISTORY ROUNDS>, you chose <PLAYER ACTION> and the opponent chose <OPPONENT ACTION>. Thus, you earned <PLAYER POINTS> points and the opponent earned <OPPONENT POINTS> points.

In total from rounds <CURRENT ROUND – HISTORY> to <PREVIOUS ROUND>, you chose J <PLAYER J HISTORY> times, Q <PLAYER Q HISTORY> times, and Z <PLAYER Z HISTORY> times. The opponent chose J <OPPONENT J HISTORY> times, Q <OPPONENT Q HISTORY> times, and Z <OPPONENT Z HISTORY> times.

You are currently playing round <CURRENT ROUND> of 40.

What do you choose for this round?

J

Q

Z

The SCoT condition also consisted the same text, consisting the baseline prompt but asking for the model's prediction of the opponents action for each round. Additions to the original prompt are highlighted in red/bold text:

You are playing a repeated game of {C.NUM_ROUNDS} rounds against the same opponent, whose decisions

will be actively made by a HUMAN participant in real time, and NOT by a computer or artificial player.

In each round, both players will make a single choice corresponding to the rules of the game described below.

Your aim is to maximise the amount of points scored which will be converted into a bonus payment at the end of the game.

The rules of the game are as follows:

Each player must select a single letter: J, Q or Z where:

J beats Z and loses to Q

Q beats J and loses to Z

Z beats Q and loses to J

In each round, the winner gets +1 point, and the loser gets -1 point.

In the event of a tie, both players earn 0 points.

You MUST choose either J, Q, or Z.

Your task is to firstly predict the next choice of the opponent in this round.

Below is the game history for the previous <HISTORY> rounds:

In round <HISTORY ROUNDS>, you chose <PLAYER ACTION> and the opponent chose <OPPONENT ACTION>. Thus, you earned <PLAYER POINTS> points and the opponent earned <OPPONENT POINTS> points.

In total from rounds <CURRENT ROUND – HISTORY> to <PREVIOUS ROUND>, you chose J <PLAYER J HISTORY> times, Q <PLAYER Q HISTORY> times, and Z <PLAYER Z HISTORY> times. The opponent chose J <OPPONENT J HISTORY> times, Q <OPPONENT Q HISTORY> times, and Z <OPPONENT Z HISTORY> times.

You are currently playing round <CURRENT ROUND> of 40.

What do you predict your HUMAN opponent will choose?

J

Q

Z

In both tasks, actual choices were forced for the SCoT condition to select the option which will beat the prediction made by the LLM. This follows evidence suggesting a distinction between what LLMs think they should choose and their actual choice, in this case between a functional and literal theory-of-mind [1].

Computational modeling

Inspection game

Parameter recovery

To validate that model parameters were identifiable from behavioural data, we conducted parameter recovery for each group's winning model. For each group we drew a single sample from the fitted posterior and took every participant's parameter values at that draw. We then simulated one synthetic choice sequence per participant, holding the employer's action at its observed values and updating the model using its own simulated choices. We then refitted the same hierarchical model to the synthetic dataset (using 3 independent chains, 30,000 burn-in and 1,000 retained samples per chain thinned by 10), and computed per-subject Pearson correlations between the true (generating) and recovered (posterior-mean) parameter values.

The first-order learning rate η recovered well across groups (r = 0.53–0.84). For the influence-play winners, the second-order weight κ and inverse temperature β also recovered well (human group, η = 0.83, κ = 0.89, β = 0.94; Gemini-SCoT, η = 0.76, κ = 0.70, β = 0.84; GPT-5, η = 0.53, κ = 0.66, β = 0.61). For GPT-4.1-SCoT, recovery was weaker but positive (η = 0.56, κ = 0.41, β = 0.55). For the RL winners, the inverse temperature β was not well recovered (DeepSeek β = −0.07; GPT-4.1 β =

0.22), consistent with its poor generative fit. Off-diagonal correlations between recovered parameters were generally small, with η, κ and β cleanly identified in the groups where the influence model won (Fig S1).

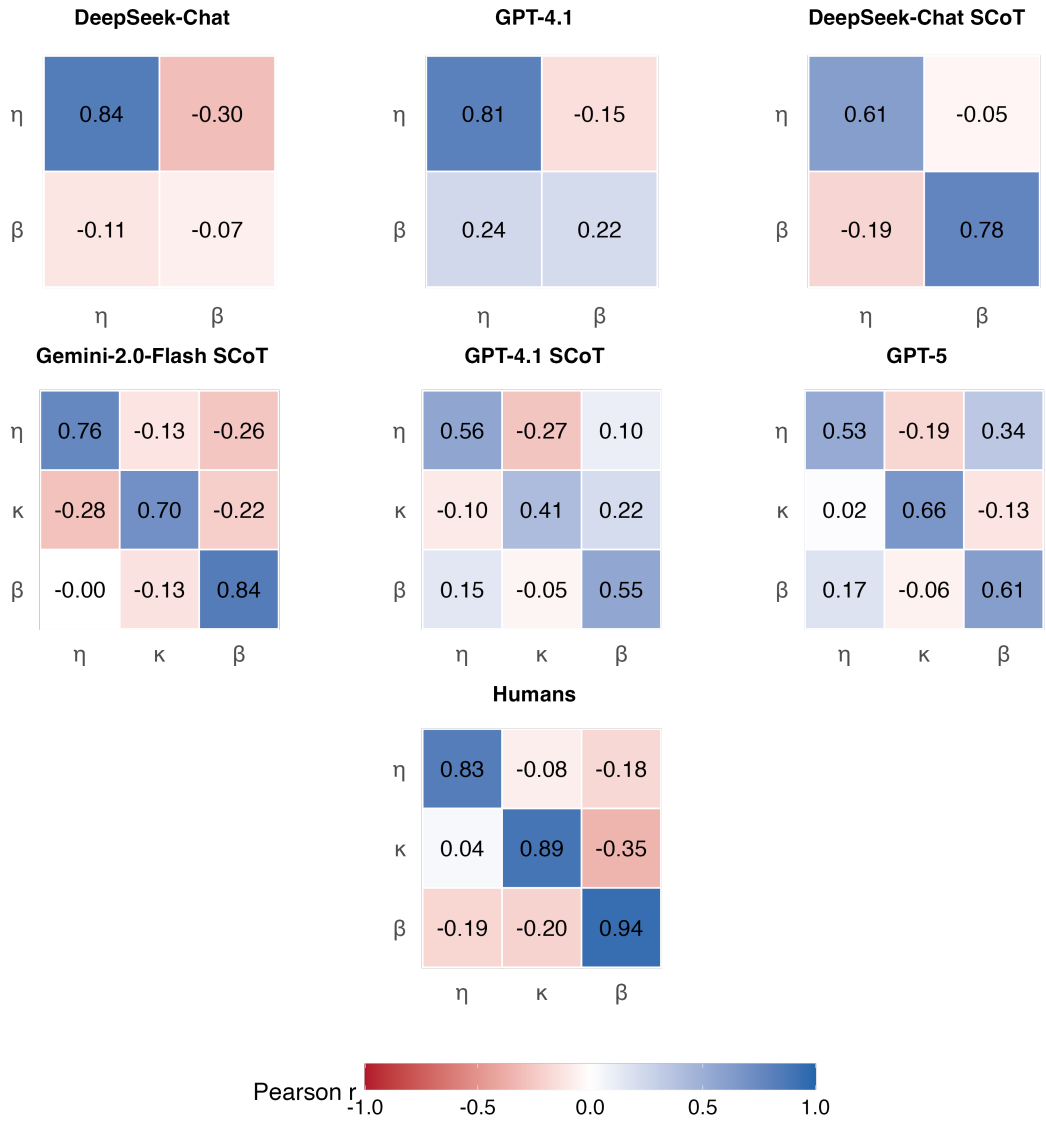


Figure S1: **Parameter recovery for the winning model in each group for the inspection game.** Each panel shows per-subject true (generated) against recovered (posterior mean) parameter values for the group's winning model.

Rock-paper-scissors

Parameter recovery

To validate that model parameters were identifiable from the behavioural data, we performed parameter recovery for each dataset using its winning model. For each subject, we extracted their fitted parameters and used them to simulate synthetic datasets. For models with a kappa (κ) parameter, we generated simulations across all possible sophistication levels ($\kappa \in \{0,1,2,3\}$), with each subject's other parameters (α, β, λ, γ) held constant at their fitted values. This approach yielded four simulated datasets per subject (one per κ level). For models without κ (i.e., 'Fictitious Play' for the prompted DeepSeek), a single simulation per subject was generated using all fitted parameters. Simulations implemented the same task structure as the original data: three blocks of 40 trials each against bot opponents of levels $k \in \{0,1,2\}$. Each simulated dataset was then fitted using the identical maximum likelihood estimation

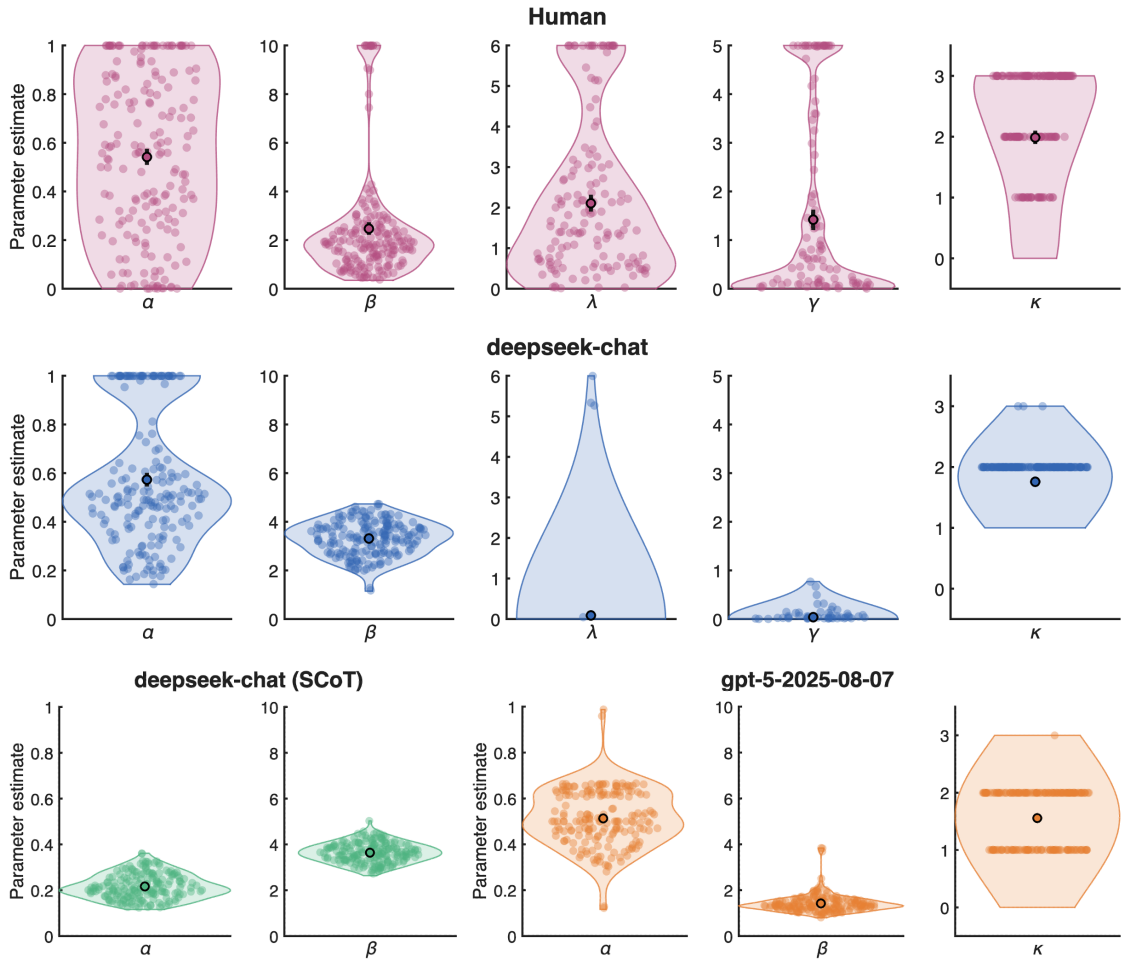


Figure S2: **Parameter estimates for the winning model in each subject group for the RPS task.** Maximum-likelihood parameter estimates reveal group and individual variability in mentalization. Note that human and DeepSeek agents are fitted to the CHASE model, DeepSeek-SCoT to the 'Fictitious Play' model and GPT-5 to the 'Theory-of-mind k' (ToMk) model. Each model has their corresponding parameters plotted: CHASE (α = speed of updating attractions, β = inverse temperature, λ = loss aversion, γ = sensitivity to evidence for opponent's level, κ = depth of mentalization ability), Fictitious Play (α = learning rate for opponent action frequencies, β = inverse temperature), ToMk (α = learning rate for confidence updates, β = inverse temperature, κ = maximum depth of recursive reasoning). Note that shared parameter letters across models may represent different cognitive processes.

procedure applied to the real data. Parameter recovery quality was assessed by computing Pearson correlations between generated and recovered parameter values across all simulations. For parameters subject to censoring (γ when $\kappa < 2$, λ when $\kappa = 0$), correlations were computed only on valid data points where the parameter was defined. Given these values, the models report good-to-excellent recovery for all winning models (all r $>$ 0.51; Fig S3).

Posterior predictive check

To validate that fitted models produced realistic behavioural patterns, we performed posterior predictive checks, where - using the same simulations generated for parameter recovery - we examined whether the models could reproduce the characteristic behavioural signatures of strategic reasoning at different opponent sophistication levels. Specifically, for each trial, we computed the best-response (BR) action to three reference strategies: (1) BR to a level-0 opponent (exploiting opponent's previous action frequencies), (2) BR to a level-1 opponent (exploiting the opponent's exploitation of the subject's frequencies), and (3) BR to a level-2 opponent (one level higher). These reference strategies represent the "signatures" of playing at sophistication levels k=1, k=2, and k=3, respectively. The frequency with which participants' actual choices matched each BR strategy ultimately provides a

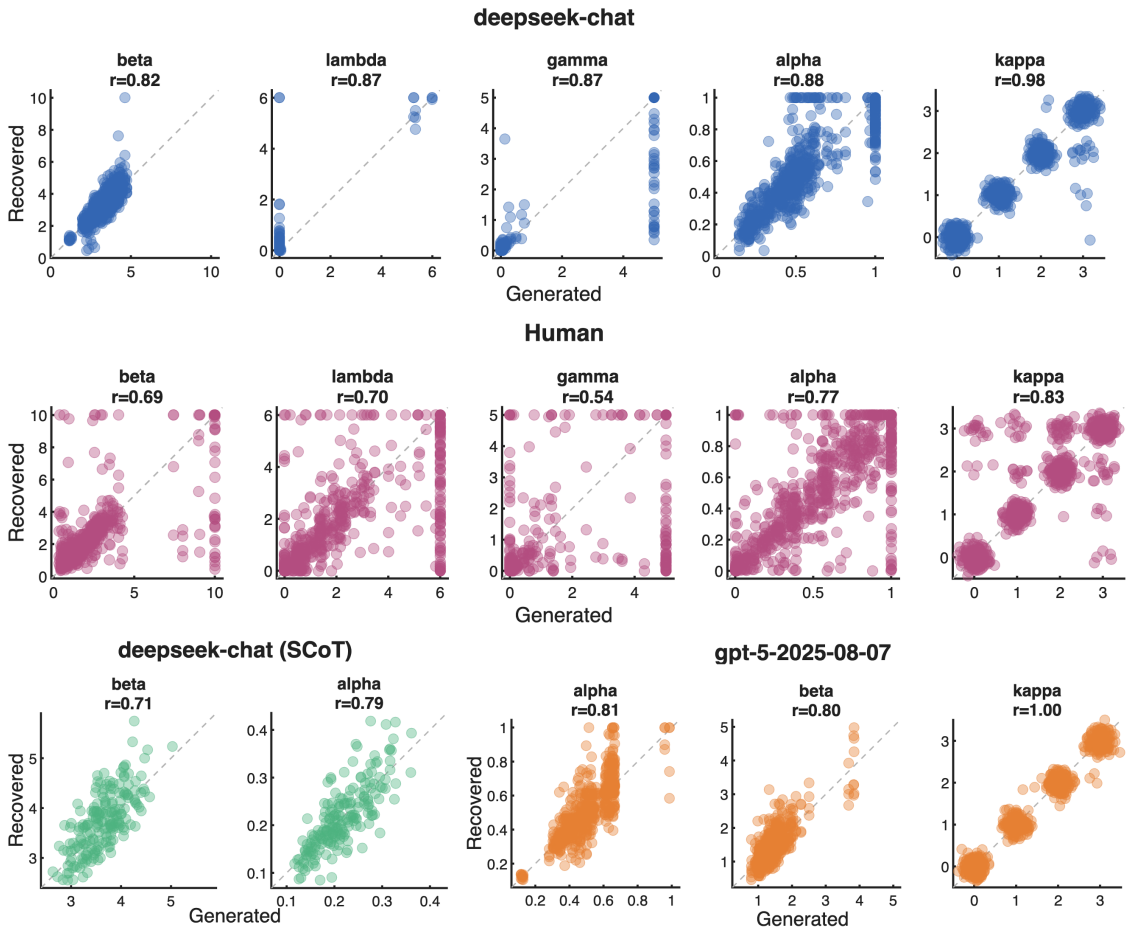


Figure S3: **Parameter recovery for the winning model in each subject group for the RPS task.** Both DeepSeek and human dataset parameters were recovered from the CHASE model, GPT-5 parameters from the ToMk model, and DeepSeek-SCoT from the Fictitious play model.

behavioural marker of their reasoning level. We then computed the proportion of trials matching each BR strategy separately for each opponent level (bot $k \in \{0,1,2\}$), using a sliding window of 5 trials. All proportions were expressed relative to chance (1/3), such that values above zero indicated above-chance usage of that strategy. This was calculated for each model and dataset combination for both real data and simulated data, along with 95% confidence intervals derived from the standard error across participants (Fig S4).

Permutation analysis of LLM participants

To validate that our modeling results were not driven by a specific combination of individual LLM instances, we performed a permutation analysis, generating 20 reshuffled datasets per condition by randomly reassigning LLM instances (blocks) to participants. Each permutation generated novel LLM participants where each subject faced all three opponent levels.

The permutation analysis revealed stable model preferences across LLM datasets, with all permutations selecting the same winning model, determined through PXP. For DeepSeek (CHASE, PXP = 1.00) and GPT-5 (ToMk, PXP = 1.00), the PXP value did not change, with the exception of DeepSeek-SCoT which changed insignificantly (Fictitious play, PXP mean = 0.96 $\pm$ 0.07). Model fit also minimally changed across reshuffles. For DeepSeek, CHASE achieved mean negLL = 100.10 $\pm$ 0.13 (SD), mean AIC = 210.21 $\pm$ 0.26, and mean BIC = 224.15 $\pm$ 0.26 across the 20 shuffles. For DeepSeek-SCoT, 'Fictitious play' showed mean negLL = 90.06 $\pm$ 0.13, mean AIC = 184.11 $\pm$ 0.25, and mean BIC = 189.69 $\pm$ 0.25. For GPT-5 'ToMk' exhibited mean negLL = 104.03 $\pm$ 0.13, mean AIC = 214.05 $\pm$ 0.26, and mean BIC = 222.42 $\pm$ 0.26.

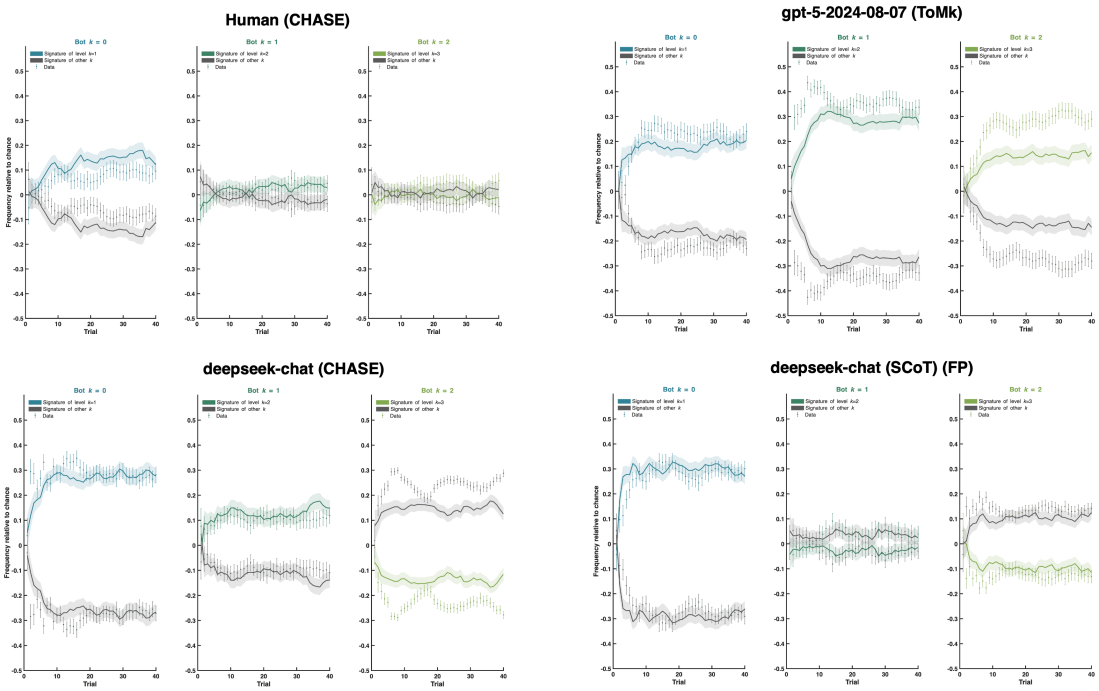


Figure S4: **Posterior predictive checks for the RPS task.** Simulated data was generated for each dataset using the recovered parameters from the respective winning model. ToMk = Theory-of-mind with confidence, FP = Fictitious Play.

Importantly, parameter estimates for winning models generally demonstrated high stability across the 20 reshuffles. For DeepSeek, CHASE model parameters showed coefficients of variation (CV) below 2% across shuffles for inverse temperature β (mean = 3.32, CV = 0.9%), learning rate α (mean = 0.56, CV = 1.7%), and theory-of-mind level κ (mean = 1.77, CV = 1.6%). Only the decay parameter γ (CV = 13.3%) and mixing parameter λ (CV = 53.1%) showed greater variability across shuffles. For DeepSeek-SCoT, parameters for the Fictitious play model were exceptionally stable across shuffles, namely for the parameters β (mean = 3.63, CV = 0.2%) and α (mean = 0.22, CV = 0.4%). Similarly, GPT-5 exhibited stable parameters across shuffles for the ToMk model, α (mean = 0.50, CV = 0.9%), β (mean = 1.45, CV = 1.4%), and κ (mean = 1.55, CV = 1.7%).

Supplementary analyses

Human/LLM analyses

Inspection game

Proportion of choices

In addition to our primary analyses, we also performed secondary analyses across human and LLM behavioural data. In the inspection game, we firstly determined the tendency for participants to select the 'work' option, where Shapiro-Wilk tests revealed significant departures from normality for Gemini, DeepSeek, and GPT-4.1 (all $p < 0.05$). For these groups, both Welch's t-tests and Mann-Whitney U tests were reported as sensitivity checks, and all other comparisons used Welch's independent samples t-tests. Bonferroni correction was applied across seven comparisons ($\alpha = 0.007$).

The majority of LLM × prompting strategy combinations showed significantly different work rates from humans (all $p < 0.001$, corrected) (Fig S5 a). Specifically, Gemini worked significantly more than humans ($t_{(113.65)} = 31.5$, $p = 3.83\times10^{-57}$, $g = 5.56$, 95% CI [4.75, 6.37]; $W = 3283$, $p = 1.37\times10^{-19}$, rank-biserial = 1.000) while DeepSeek ($t_{(84.57)} = -28.8$, $p = 1.36\times10^{-44}$, $g = -4.75$, 95% CI [-5.46, -4.03]; $W = 1.0$, $p = 5.58\times10^{-20}$, rank-biserial = -0.999) and GPT-4.1 ($t_{(84.41)} = -25.5$, $p = 1.33\times10^{-40}$, $g = -4.21$, 95% CI [-4.86, -3.55]; $W = 2.5$, $p = 9.26\times10^{-20}$, rank-biserial = -0.999) worked significantly less.

Among the remaining comparisons, Gemini (SCoT) ($t_{(82.15)}$ = -9.26, p = 1.52×10^{-13}, g = -1.52, 95% CI [-1.93, -1.11]) worked less whilst GPT-4.1 (SCoT) ($t_{(75.76)}$ = 12.28, p = 7.53×10^{-19}, g = 2.00, 95% CI [1.55, 2.44]) and DeepSeek (SCoT) ($t_{(79.11)}$ = 5.54, p = 2.74×10^{-6}, g = 0.91, 95% CI [0.52, 1.29]) worked more. The only exception was GPT-5, which did not significantly differ from humans ($t_{(74.79)}$ = 0.70, p = 0.486, g = 0.13, 95% CI [-0.18, 0.44]).

We then compared work rates across prompting conditions within each model. Due to normality violations in the normal conditions, both Welch's t-tests and Mann-Whitney U tests were reported for all three models. Bonferroni correction was applied across three comparisons (α = 0.017). All three models showed significantly different work rates between strategies (all p < 0.001, corrected). Specifically, DeepSeek ($t_{(94.79)}$ = -73.09, p = 1.10×10^{-84}, g = -14.51, 95% CI [-16.54, -12.40]; W = 0, p = 1.06×10^{-17}, rank-biserial = -1.000) and GPT-4.1 ($t_{(88.77)}$ = -84.88, p = 2.66×10^{-86}, g = -16.85, 95% CI [-19.19, -14.41]; W = 0, p = 1.85×10^{-17}, rank-biserial = -1.000) worked more in SCoT than the baseline condition whilst Gemini worked more in the baseline condition than SCoT ($t_{(66.25)}$ = 57.05, p = 1.42×10^{-57}, g = 11.46, 95% CI [9.77, 13.09]; W = 2450, p = 2.95×10^{-17}, rank-biserial = 1.000). Examining the proportion of participants selecting to work on each trial further highlighted differences in work rate across SCoT and normally prompted LLMs (Fig S5 c). As with the group-level results, DeepSeek and GPT-4.1 demonstrated relative homogeneity across participants without strategic prompting, which became more diverse under SCoT. However, this pattern was reversed for the Gemini model.

First trial choices

Large language models demonstrate biases in one-shot decision making [2], therefore we set to similarly assess the probability of choosing to work specifically on the first trial (Fig S5 b). In the inspection game, LLMs were biased to select a particular choice option, with 5/7 models selecting one option (either 'work' or 'shirk') across all participants. Humans, on the other hand, selected with a slight bias against working (43.3% work, 56.7% shirk). However, non-absolute patterns in choice were observed for DeepSeek, (p_{Work} = 0.68) and GPT-5 (p_{Work} = 0.57). Comparing one-shot work choices for these two groups with humans (Bonferroni-corrected α = 0.025), DeepSeek worked significantly more often (Welch's $t_{(108.79)}$ = 2.736, p = 0.015, 95% CI: [0.0681, 0.4262], Hedges' g = 0.50, 95% CI [0.13, 0.87]; sensitivity check: $\chi^2(1) = 6.08$, p = 0.027), whereas GPT-5 did not significantly differ (Welch's $t_{(141.35)}$ = 1.743, p = 0.167, 95% CI: [-0.0184, 0.2928], Hedges' g = 0.27, 95% CI [-0.04, 0.58]; sensitivity check: $\chi^2(1) = 2.50$, p = 0.228).

Rock-paper-scissors

Proportion of choices

In the RPS task we firstly examined whether choice proportions (Humans: 'Rock', 'Paper', 'Scissors'; LLMs: 'J', 'Q', 'Z') differed from random selection (33.3%) using exact binomial tests (Fig S6 a), following previous results suggesting non-random choice [2]. Choice frequencies were aggregated across all participants within each group, yielding a total of 12 comparisons (4 groups × 3 choices), using a Benjamini-Hochberg correction for the false discovery rate. We also calculated odds ratios (OR) expressing the odds of selecting an option relative to the odds expected under chance (1/3), where an odds ratio of 1 indicates chance-level selection.

Humans exhibited no significant deviations from chance across all trials (Rock: 34.0%, p = 0.071, OR = 1.03; Paper: 33.2%, p = 0.705, OR = 0.99; Scissors: 32.8%, p = 0.148, OR = 0.98, all corrected), indicating approximately uniform selection across the three options. In contrast, all LLM groups showed significant choice biases. DeepSeek demonstrated a significant preference for 'J' (35.5%, p = 3.06×10^{-11}, OR = 1.10) and bias against 'Q' (31.0%, p = 6.24×10^{-13}, OR = 0.90), while 'Z' selections remained at chance (33.5%, p = 0.678, OR = 1.01). DeepSeek-SCoT showed the opposite pattern, choosing 'J' less often (26.2%, p = 5.35×10^{-119}, OR = 0.71), and choosing 'Q' (39.4%, p = 1.58×10^{-79}, OR = 1.30) and 'Z' (34.5%, p = 7.97×10^{-4}, OR = 1.05) above chance level. GPT-5 exhibited a more subtle bias,

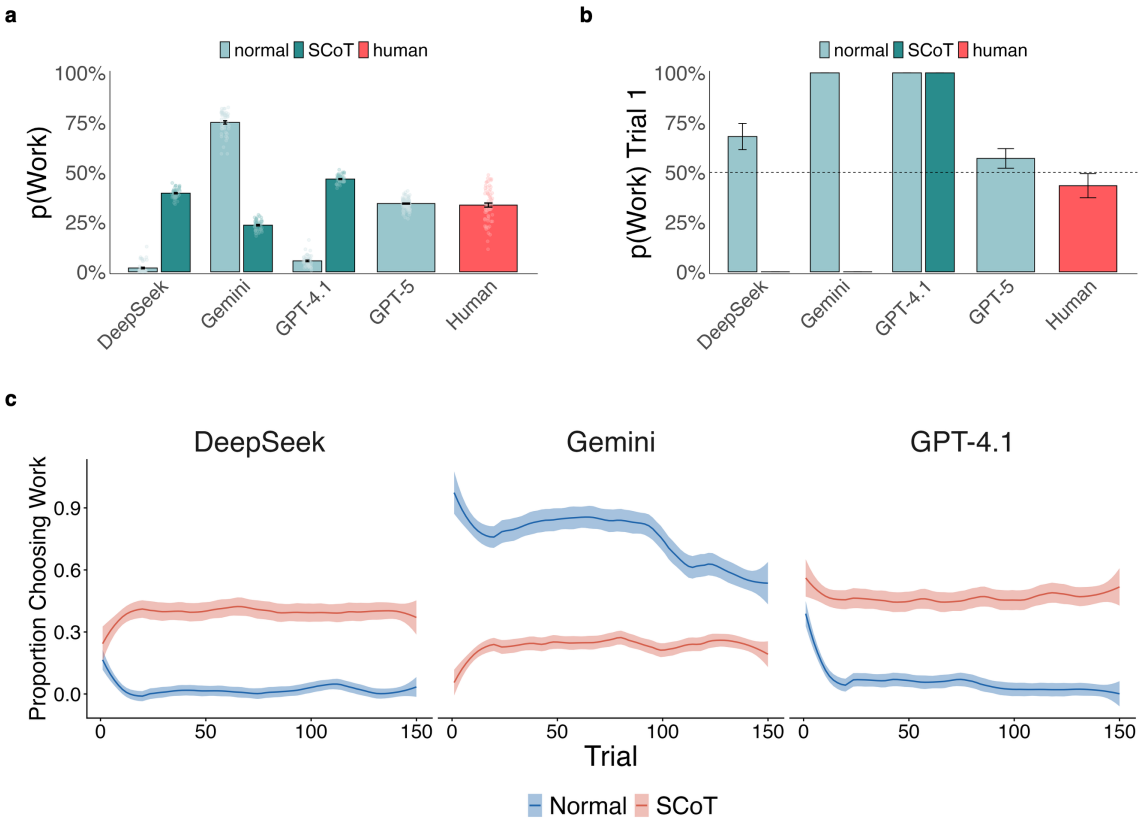


Figure S5: **Work rates among LLMs and humans by model and prompting strategy in the inspection game. (a)** Work rates across all trials averaged by subject revealed significant differences amongst LLMs and human participants. **(b)** When only considering the choice to work on the very first trial, human participants show a slight bias against working, opposite to GPT-5 which demonstrates a slight bias for working. Contrarily, most other LLMs demonstrate complete favouritism towards either selecting to work or shirk. **(c)** LLMs demonstrate different patterns of heterogeneity in the proportion of participants choosing to work across trials. For each group, we computed the proportion of participants choosing the work option on each trial, which were then smoothed using locally weighted regression (LOESS; span = 0.3).

selecting 'J' and 'Z' at chance levels (J: 32.7%, p = 0.062, OR = 0.97; Z: 32.9%, p = 0.193, OR = 0.98) only showing a positive selection bias for 'Q' (34.4%, p = 0.001, OR = 1.05).

First trial choices

Initial choice patterns for the first trial revealed more pronounced biases than overall choice probabilities, particularly for LLMs (Fig S6 b). Again, binomial tests against chance (33.3%) were conducted for each group × choice combination (12 tests), with Benjamini-Hochberg corrections applied. Humans now showed an initial Rock bias (41.8%, p = 1.53×10^{-4}, OR = 1.44), with Paper (28.9%, p = 0.046, OR = 0.81) and Scissors (29.3%, p = 0.066, OR = 0.83) biased against, with the latter result narrowly missing significance after multiple comparison correction. Contrarily, DeepSeek showed a dominant preference for choosing 'J' (83.6%, p = 1.73×10^{-133}, OR = 10.17), with a bias against selecting 'Q' (16.4%, p = 2.28×10^{-19}, OR = 0.39) and not selecting 'Z' at all. GPT-5 now similarly exhibited a strong 'Q' bias (82.3%, p = 9.39×10^{-127}, OR = 9.32), with 'J' (6.5%, p = 8.76×10^{-53}, OR = 0.14) and 'Z' (11.1%, p = 2.66×10^{-34}, OR = 0.25) both biased against. Interestingly, DeepSeek-SCoT displayed complete determinism, with all 568 participants selecting 'Q' on trial 1.

Adaptive mentalization (winner-take-all)

As a complementary analysis to restricting trials only with a dominant belief, we also examined adaptive mentalization using a winner-take-all approach, where the opponent level with the highest belief probability on each trial was designated as the inferred level. This analysis subsequently retains all

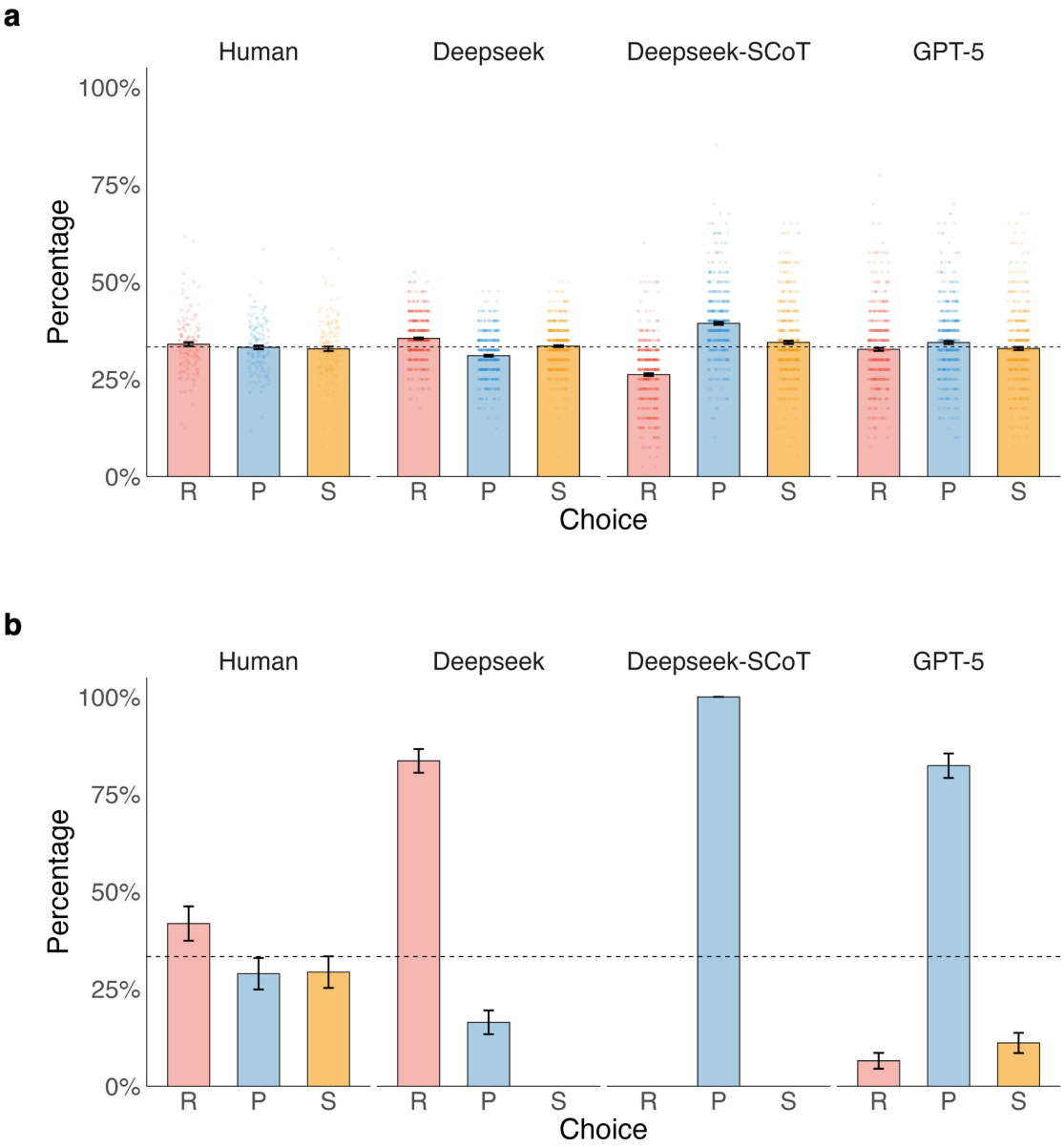


Figure S6: **Choice proportions among LLMs by model and prompting strategy in the rock-paper-scissors game. (a)** Choice frequencies across all trials averaged by subject revealed significant differences amongst LLMs and human participants, where humans demonstrate no significant bias for any choice option, a pattern not observed in any LLM. **(b)** When only considering the choice for the very first trial, human participants show a slight bias towards selecting 'rock', a pattern differing significantly among LLMs. In both plots, choice options have been recoded to follow the standard 'RPS' format. In the tasks, options were presented as 'Rock', 'Paper', 'Scissors' for humans and 'J', 'Q', 'Z' for LLMs.

modelled trials. In cases where multiple beliefs were tied, one was selected at random. Results revealed substantial heterogeneity across populations (Fig S7). GPT-5 agents were able to accurately adjust across all opponent levels (k=0: 87.95%, k=1: 83.99%, k=2: 79.92%; overall: 83.95% correct) with 95.2% of participants inferring the correct opponent level on more than half of all trials, indicating adaptive mentalization. Furthermore, ties were relatively rare in this group, occurring in only 6.29% of trials. Conversely, DeepSeek agents, whilst able to accurately identify and adapt to non-strategic opponents (k=0: 72.89% correct), performed at chance level against higher-reasoning opponents (k=1: 48.42%, k=2: 0.36% correct). Notably, nearly half of all trials (49.16%) resulted in tied beliefs, primarily among participants with $\kappa \geq 2$ (66.45% of $\kappa=2$ trials, 68.33% of $\kappa=3$ trials), suggesting uncertainty. Human participants exhibited intermediate performance (k=0: 54.95%, k=1: 53.79%, k=2: 26.33%

correct), with 40.5% of participants inferring the correct opponent level on more than half of all trials and 17.41% of trials showing tied beliefs. Critically, this approach revealed the same qualitative patterns as the primary analysis, where GPT-5 agents successfully inferred opponent sophistication at all levels, DeepSeek agents correctly identified non-strategic opponents but failed to distinguish between strategic levels, and humans showed above-chance inference for k=0 and k=1 opponents but struggled with k=2 opponents.

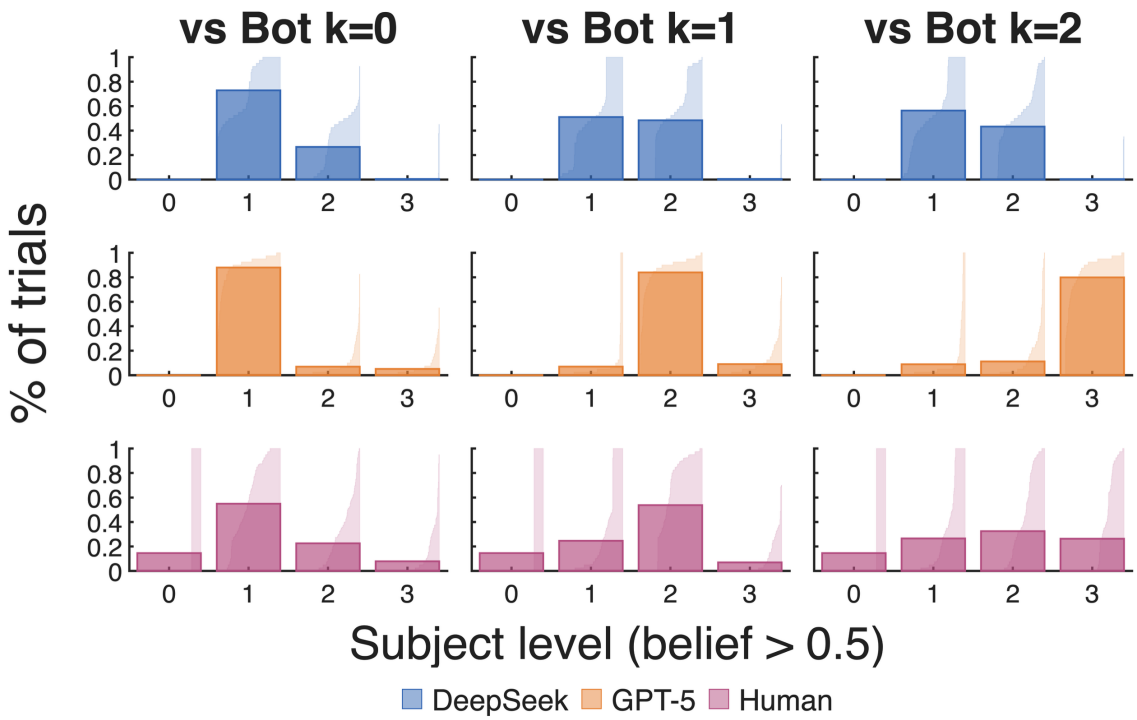


Figure S7: **The inferred level of mentalization among agents using the winner-takes-all approach.** Reflecting the primary approach where trials with a dominant belief (>0.5) were included, DeepSeek and human participants adapted well to low-adaptive reasoners, but struggled with opponents capable of higher-order mentalization. Conversely, GPT-5 participants were able to adapt well to all reasoning levels. DeepSeek-SCoT participants were again excluded from the plot, being almost exclusively estimated to have a value of k<2 when fit with the CHASE model.

Checks

Human opponent

Both tasks, when conventionally used among human participants, involve instructing or convincing the subject that they are playing against another human. Consequently, the subject is implicitly guided to engage in mentalization, which they may not be inclined to do believing they were playing against an algorithm. Therefore, it was vital that the model also believed that they were playing against a human opponent and not a computerized agent.

Subsequently, for the Inspection game on each trial, in addition to the task information, the model was also asked whether it believed to be playing against a human or computerised opponent:

<GAME HISTORY>

<GAME RULES>

<CHOICE QUESTION>

Who do you think you are playing against currently, a HUMAN or a COMPUTER?

HUMAN

COMPUTER

All models reported that they believed to be playing against a human opponent in all trials, and explicitly referenced the specific instruction. This question was subsequently not included in the RPS game implying the same reasoning across tasks, and considering monetary costs.

Comprehension

To rule out the confound where behavioural differences may instead be explained through issues with understanding the task rules and recalling the choice history, a series of comprehension questions adapted from a previous study [3] were used to determine whether the models were able to accurately retrieve important information regarding the game history and rules.

In the comprehension assessment, LLMs completed the same experimental task with several questions presented at the end of each trial:

Inspection game

1. In cents, what is the {lowest or highest} payoff the EMPLOYEE can earn in a single round?
2. In cents, what is the {EMPLOYEE or EMPLOYER} payoff in a single round if {EMPLOYEE/EMPLOYER} chooses to {ACTION} and {EMPLOYEE/EMPLOYER} chooses to {ACTION}?
3. What is the current round of the game?
4. Which action did the {EMPLOYEE/EMPLOYER} choose in round {X}?
5. How many cents did the {EMPLOYEE/EMPLOYER} earn in round {X}?
6. How many times has the {EMPLOYEE/EMPLOYER} chosen to {ACTION} within the past {N} rounds?
7. If the EMPLOYER decides to {INSPECT or NOT INSPECT}, which action should the EMPLOYEE choose to maximise their payoff, WORK or SHIRK?

Rock-paper-scissors

1. What is the {lowest or highest} number of points a player can earn in a single round?
2. How many points do {YOU/YOUR OPPONENT} earn in a single round if you choose {J/Q/Z} and your opponent chooses {J/Q/Z}?
3. What is the current round of the game?
4. Which choice did {YOU/YOUR OPPONENT} make in round {X}?
5. How many points did {YOU/YOUR OPPONENT} earn in round {X}?
6. How many times did {YOU/YOUR OPPONENT} choose {J/Q/Z} within the past {N} rounds?
7. If your opponent chooses {J/Q/Z}, which choice should you make to maximize your points?

Questions 1, 2, 3 and 7 were designed to test the models' understanding of the game rules and appropriate actions, whilst questions 4, 5 and 6 were designed to test the models' understanding of previous actions provided within the game history. In each trial, where appropriate, questions independently sample at random a value for a trial (Q4, Q5) and a range of trials (Q6) within the defined history. Where appropriate, roles are also determined randomly on each trial. The questions were presented for each trial (with the exception of the very first trial) included in analyses. For the inspection game, 3 sessions were run for each model, resulting in 447 trials. For the RPS game, 5 sessions were run for each model, resulting in 195 trials. The choice history of 10 trials kept unchanged from the primary analyses.

Results demonstrate that across both tasks larger models were able to accurately recall the game history and game rules consistently, whilst smaller models were less accurate (Table S1 and Table S2). However, models were generally less accurate on RPS compared to the inspection game. These results subsequently informed the models kept for the primary analyses.

Table S1: Comprehension accuracy for Large Language Models in the inspection game. Accuracy is calculated across 447 trials (3 sessions × 150 trials, not including the first trial). Included models in the primary set of analyses are highlighted in bold.

model	time_accuracy	choice_frequency_accuracy	strategic_accuracy	points_accuracy	payoff_accuracy	action_accuracy	rules_accuracy	overall
claude-3-5-sonnet-20241022	1	1	1	1	1	1	1	1
claude-3-7-sonnet-20250219	1	1	1	1	1	1	1	1
claude-opus-4-1-20250805	1	1	1	1	1	1	1	1
claude-opus-4-20250514	1	1	1	1	1	1	1	1
claude-sonnet-4-20250514	1	1	1	1	1	1	1	1
claude-sonnet-4-5-20250929	1	1	1	1	1	1	1	1
gemini-1.5-pro	1	1	1	1	1	1	1	1
gpt-4-turbo-2024-04-09	1	1	1	1	1	1	1	1
gpt-4.1-2025-04-14	1	1	1	1	1	1	1	1
o3-2025-04-16	1	1	1	1	1	1	1	1
gpt-4.1-mini-2025-04-14	1	1	1	1	1	0.998	1	1.000
gpt-5-2025-08-07	1	1	1	1	1	0.998	1	1.000
deepseek-chat	1	1	0.996	1	1	1	1	0.999
gemini-2.0-flash	1	1	1	0.996	1	1	1	0.999
claude-3-opus-20240229	1	1	1	1	0.969	1	1	0.996
claude-3-5-haiku-20241022	1	1	0.993	1	0.998	0.960	0.975	0.989
gpt-5-mini-2025-08-07	1	1	1	1	1	0.919	1	0.988
llama-4-maverick-17b-128e-instruct	1	1	1	1	0.996	0.658	1	0.950
llama-3.3-70b-versatile	1	0.998	1	1	1	0.530	1	0.933
gpt-5-nano-2025-08-07	1	1	0.998	0.998	0.971	0.553	0.991	0.930
gpt-4o-mini-2024-07-18	1	1	0.700	0.906	0.913	0.490	1	0.858
gpt-4.1-nano-2025-04-14	1	0.801	0.530	0.924	0.702	0.539	0.998	0.785

Table S2: Comprehension accuracy for Large Language Models in the rock-paper-scissors game. Accuracy is calculated across 195 trials (5 sessions × 40 trials, not including the first trial). Included models in the primary set of analyses are highlighted in bold.

model	time_accuracy	choice_frequency_accuracy	strategic_accuracy	points_accuracy	payoff_accuracy	action_accuracy	rules_accuracy	overall_accuracy
gemini-1.5-pro	1	0.995	1	1	0.990	1	1	0.998
o3-2025-04-16	1	1	1	1	1	0.979	1	0.997
gpt-5-2025-08-07	1	1	1	1	1	0.969	1	0.996
claude-opus-4-1-20250805	1	1	0.969	1	0.969	1	1	0.991
claude-3-5-sonnet-20241022	1	1	1	1	0.933	1	1	0.990
claude-opus-4-20250514	1	1	0.974	1	0.959	1	1	0.990
gpt-4o-2024-08-06	1	1	0.974	1	0.954	1	1	0.990
claude-3-7-sonnet-20250219	1	1	0.995	1	0.913	1	1	0.987
gpt-4.1-2025-04-14	1	1	0.933	1	0.979	0.954	1	0.981
claude-sonnet-4-5-20250929	1	1	0.979	1	0.908	0.979	1	0.981
gemini-2.0-flash	1	1	0.908	1	0.938	1	1	0.978
claude-sonnet-4-20250514	1	1	0.954	1	0.892	0.990	1	0.977
deepseek-chat	1	1	0.856	1	0.918	1	1	0.968
gpt-5-mini-2025-08-07	1	0.995	1	1	0.995	0.774	1	0.966
gpt-4-turbo-2024-04-09	1	1	0.979	1	0.923	0.841	1	0.963
claude-3-opus-20240229	1	1	0.692	1	0.877	0.995	1	0.938
gpt-4.1-mini-2025-04-14	1	1	0.779	1	0.918	0.646	1	0.906
llama-4-maverick-17b-128e-instruct	1	0.990	0.708	0.995	0.815	0.703	0.995	0.886
gpt-4o-mini-2024-07-18	1	1	0.8	0.985	0.708	0.538	1	0.862
claude-3-5-haiku-20241022	1	1	0.523	1	0.8	0.672	1	0.856
llama-3.3-70b-versatile	1	0.974	0.631	0.995	0.759	0.390	1	0.821
gpt-4.1-nano-2025-04-14	1	0.892	0.426	0.754	0.621	0.528	1	0.746
gpt-5-nano-2025-08-07	0.697	0.646	0.605	0.636	0.646	0.446	0.656	0.619

Context manipulation

Large language models are ultimately trained upon large quantities of text-based data. Therefore, LLM responses may be influenced by subtle phrasing in the prompt, perception of roles and order effects. Whilst the construction of the experimental prompts used were chosen to reflect the information provided to human participants, we also manipulated our prompts to assess these biases.

We specifically focus on the following two biases:

1. **Order.** Large language models may be biased to simply select the choice option presented first in the sequence.
2. **Language.** Models may simply be more likely to select the choice option more frequently appearing within the training data, or be biased towards options with certain meaning, i.e., 'work' instead of 'shirk', as working is a positive action more likely to appease the user.

To investigate these effects in the inspection game, we simulated sessions for recoded choice options. Only the baseline prompting scheme was tested under these conditions. Choice options for the employee were recoded to single letters 'J' and 'X', infrequent letters within the English corpora which feature with a roughly equal probability. Their position was reversed in a second condition, presenting the choice options as the single letters 'X' and 'J'. Across both conditions, the experimental roles of 'EMPLOYEE' / 'EMPLOYER' were replaced with 'You' / 'Opponent' to remove potential attribution biases. The choice options for the 'Employer' were replaced with the single letters 'Q' and 'Z' in both conditions. The order of the choice options within the prompt was 'X' and 'J' for both conditions.

The prompt for with the 'X / J' re-coding of options is presented below:

> Round <CURRENT ROUND> of 150
>
> You are participating in a game where the goal of both players is to maximize their respective earnings.
>
> In the game, there are two players, you and the opponent. You will play 150 repeated rounds with the same opponent, whose decisions will be actively made by a HUMAN participant in real time, and NOT by a computer or artificial player.
>
> In each round the opponent will have to make a choice, whether to choose Q or Z. In each round you will have to make a choice, whether to choose X or J.
>
> The rules of the game are as follows:
>
> If the opponent chooses Z and you choose X, the opponent earns 100 cents, and you earn 0 cents.
>
> If the opponent chooses Q and you choose X, the opponent earns 0 cents, and you earn 50 cents.
>
> If the opponent chooses Z and you choose J, the opponent earns 0 cents and you earn 50 cents.
>
> If the opponent chooses Q and you choose J, the opponent earns 25 cents, and you earn 0 cents.
>
> Below is the game history for the previous <HISTORY> rounds:
>
> <GAME HISTORY>
>
> In total from rounds <HISTORY>:
>
> The OPPONENT chose Q <N> times and chose Z <N> times.
>
> YOU chose X <N> times and chose J <N> times.
>
> You are currently playing round <CURRENT ROUND>.
>
> What do you choose?
>
> X
>
> J

For the rock-paper-scissors game, the original choice options 'J' / 'Q' / 'Z' were replaced with 'Q' / 'Z' / 'J' and 'Z' / 'J' / 'Q'. Otherwise, the prompts were left unchanged. The choice options within the prompt were also re-ordered to match the recoding sequence.

For example, the prompt for the 'Z / J / Q' re-coding of options is presented below:

The rules of the game are as follows:

Each player must select a single letter: Z, J or Q where:

Z beats J and loses to Q

J beats Q and loses to Z

Q beats Z and loses to J

In each round, the winner gets +1 point, and the loser gets -1 point.

In the event of a tie, both players earn 0 points.

You MUST choose either Z, J or Q.

Below is the game history for the previous <HISTORY> rounds:

<GAME HISTORY>

In total from rounds <HISTORY>:

You chose Z <N> times, J <N> times, and Q <N> times.

The opponent chose Z <N> times, J <N> times, and Q <N> times.

You are currently playing <CURRENT ROUND> of 40.

What do you choose for this round?

Z

J

Q

As exploratory analyses, formal hypothesis tests were not planned. Given time and financial constraints, ~25 sessions were run in each condition for the inspection game, and 30 sessions for the rock-paper-scissors game for each programmed k-level of the opponent. As the recoded prompt did not incorporate strategic prompting, each group was then compared with an equal-size randomly sampled group of participants responding to the choice options as originally coded.

Inspection game

Mean Payoff

We compared the mean payoff per trial across the three paradigm conditions (work/shirk vs. X/J vs. J/X) within each model using Welch's one-way ANOVAs, as sample sizes varied across paradigms (work/shirk: n = 49-100; X/J: n = 25; J/X: n = 24-27). Gemini showed significant paradigm effects (Welch's $F_{(2,\ 59.07)} = 403.636$, $p = 3.57\times10^{-35}$, f = 2.49), with Games-Howell post-hoc tests revealing all three paradigms differed significantly from one another (all $p < 0.001$). DeepSeek also demonstrated significant paradigm effects (Welch's $F_{(2,\ 52.97)} = 7.831$, $p = 0.001$, f = 0.37), where J/X recoding (M = 18.33, SD = 1.41) did not differ significantly from the original work/shirk coding ($p = 0.543$), with the comparison against the X/J recoding presenting a marginal difference ($p = 0.063$). In contrast, both GPT-5 (Welch's $F_{(2,\ 44.38)} = 3.072$, $p = 0.056$, f = 0.21) and GPT-4.1 (Welch's $F_{(2,\ 44.61)} = 0.661$, $p = 0.521$, f = 0.12) showed no significant paradigm effects, suggesting that framing of choice options did not substantially affect strategic performance for these models (Fig S8).

Rock-paper-scissors

Mean Payoff

In the RPS task, we first examined overall mean payoffs using a two-way between-participants ANOVA with LLM (DeepSeek vs. GPT-5) and choice order (JQZ vs. QJZ vs. ZJQ) as factors, collapsing across bot level. Sample sizes were filtered to be unbalanced across choice order conditions (JQZ: n = 564 per model; QJZ: n = 90 per model; ZJQ: n = 90 per model), and Levene's test indicated significant variance heterogeneity ($F_{(5,\ 1482)} = 107.73$, $p = 3.59\times10^{-97}$), with a maximum-to-minimum variance ratio of 5.71:1. Due to this heterogeneity and the unbalanced design, we used Type III sums of squares and subsequently verified conclusions with robust regression, permutation tests, and bootstrap confidence

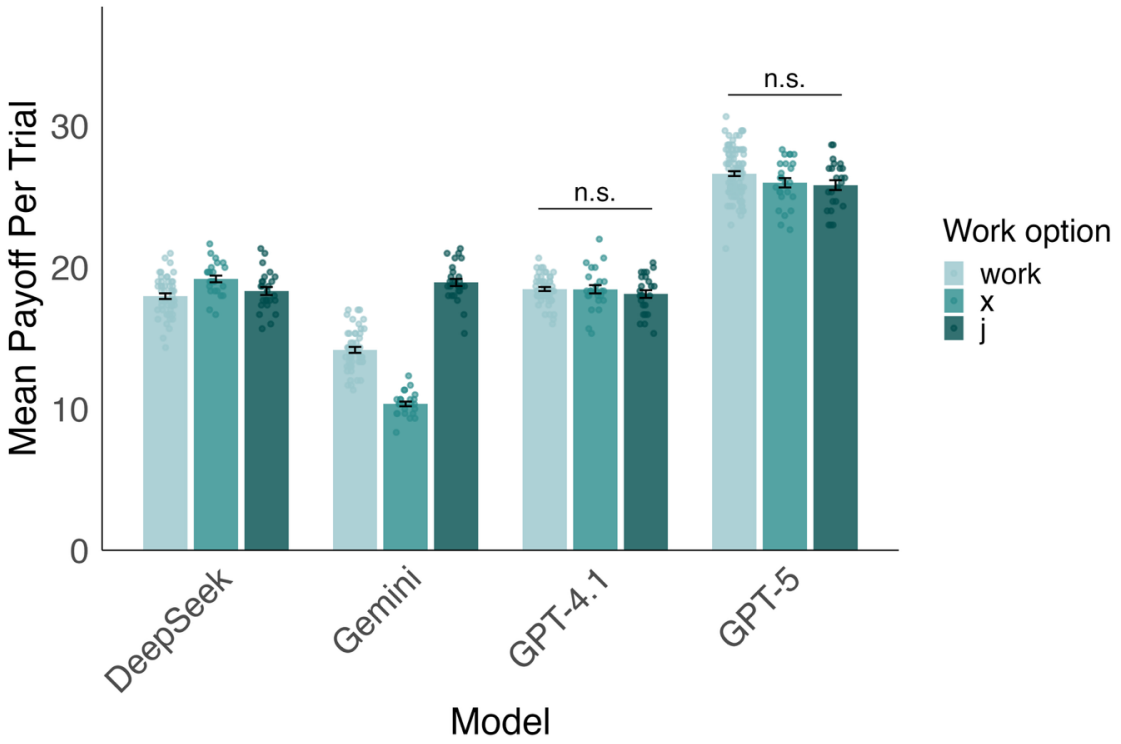


Figure S8: **Effects of order and language on performance by Large Language Models in the inspection game.** Recoding the 'work' variable significantly impacted task performance among DeepSeek and Gemini participants, but not in GPT-4.1 and GPT-5 agents.

intervals. Results revealed a large main effect of LLM ($F_{(1, 1482)} = 376.83$, $p = 5.68\times10^{-75}$, $f = 0.50$), with GPT-5 earning higher mean payoffs than DeepSeek. In contrast, there was no significant effect of choice order ($F_{(2, 1482)} = 0.46$, $p = 0.633$, $f = 0.03$), nor an interaction effect ($F_{(2, 1482)} = 0.07$, $p = 0.928$, $f = 0.01$), indicating that the presentation order of options (JQZ, QJZ, or ZJQ) did not influence performance outcomes. Tukey HSD post-hoc tests also confirmed that mean payoffs did not differ across choice orders (all pairwise $p > 0.62$).

To specifically examine whether model performance varied as a function of opponent sophistication, we conducted a three-way between-participants ANOVA with LLM (DeepSeek vs. GPT-5), choice order (JQZ vs. QJZ vs. ZJQ), and bot level (0 vs. 1 vs. 2) as factors, maintaining Type III sums of squares for the unbalanced design. Levene's test indicated severe variance heterogeneity ($F_{(17, 1470)} = 6.60$, $p = 2.16\times10^{-15}$), with an 18:1 maximum-to-minimum variance ratio, driven primarily by DeepSeek exhibiting increased variability against level-2 bots (SD increase from 0.046-0.078 at level-1 to 0.129-0.196 at level-2). We conducted similar sensitivity analyses as previously described. Results (Fig S9) revealed significant main effects of LLM ($F_{(1, 1470)} = 1402.84$, $p = 3.92\times10^{-216}$, $f = 0.98$) and bot level ($F_{(2, 1470)} = 482.14$, $p = 9.89\times10^{-162}$, $f = 0.82$), but no effect of choice order ($F_{(2, 1470)} = 1.70$, $p = 0.182$, $f = 0.05$). Critically, a large LLM × bot level interaction emerged ($F_{(2, 1470)} = 580.74$, $p = 1.35\times10^{-186}$, $f = 0.89$), confirming that model performance differences increased dramatically with opponent sophistication. None of the other two-way interactions approached significance (LLM × choice order: $F_{(2, 1470)} = 0.28$, $p = 0.758$; choice order × bot level: $F_{(4, 1470)} = 1.02$, $p = 0.395$), and the three-way interaction was marginal with a small effect size ($F_{(4, 1470)} = 2.11$, $p = 0.078$, $f = 0.08$), suggesting that choice order did not moderate the LLM x bot level interaction.

Simple effects analyses at each bot level further revealed that the improved performance in GPT-5 over DeepSeek escalated with increasing bot sophistication, repeating the pattern observed in the primary analysis. Specifically, a small but significant advantage was observed for GPT-5 at level 0 ($p = 0.002$, $d = -0.39$), a substantially larger effect at level 1 ($p = 4.75\times10^{-36}$, $d = -1.63$), and an exceptionally large difference at level 2 as DeepSeek performance worsened while GPT-5 remained performing above

chance ($p < 1\times10^{-300}$, $d = -6.18$). Sensitivity tests once again confirmed these conclusions remained valid despite variance heterogeneity (all $p < 0.001$), and bootstrap confidence intervals for model differences at each bot level all excluded zero (Bot 0: 95% CI [-0.060, -0.026]; Bot 1: 95% CI [-0.190, -0.161]; Bot 2: 95% CI [-0.650, -0.605]).

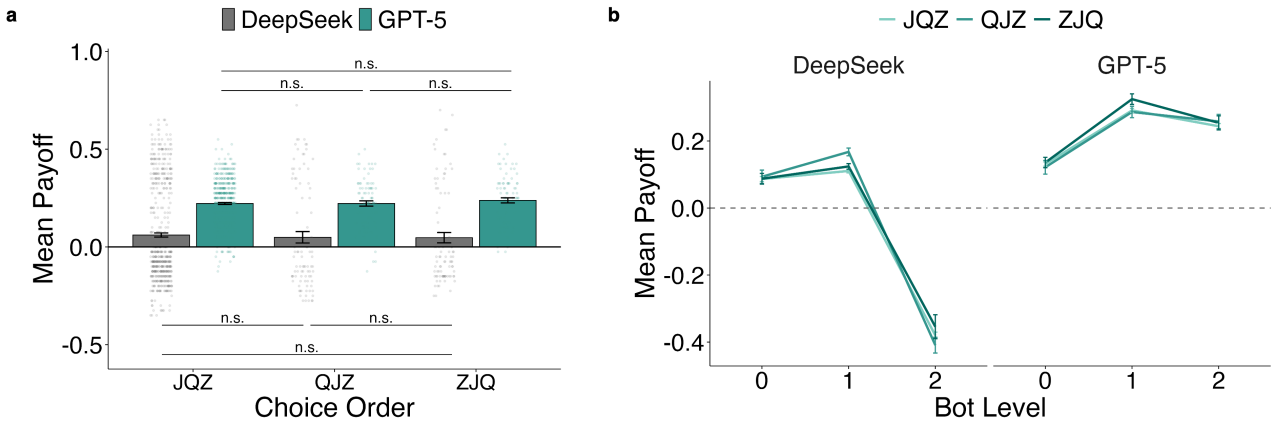


Figure S9: **Effects of order and language on task performance by Large Language Models in the rock-paper-scissors game.** Recoding the choice options did not affect performance in the task for either LLM measured through the mean payoff.

Choice Probability

As with the original codings, we similarly investigated whether models' choice distributions deviated from equal selection (33.3% for each option) across all 40 trials for the different coding schemes. For this purpose, we employed binomial tests with Benjamini-Hochberg correction for multiple comparisons (Fig S10 a), where only baseline prompting conditions were included. On the whole, DeepSeek showed strong position-dependent biases that varied systematically with choice order ($\chi^2(4, N = 29{,}760) = 457.09$, $p = 2.55\times10^{-97}$, Cramer's $V = 0.088$). Specifically, in the JQZ condition, DeepSeek slightly over-selected the first option (J: 35.5%, $p = 2.19\times10^{-11}$) and under-selected the second option (Q: 31.0%, $p = 5.24\times10^{-13}$), while selecting the third option at chance levels (Z: 33.5%, $p = 0.641$). In the QJZ condition, a stronger first-option bias emerged (Q: 41.7%, $p = 5.6\times10^{-25}$), with significant underselection of the second-position option (J: 23.4%, $p = 1.88\times10^{-37}$). In the ZJQ condition, the bias was most extreme, with DeepSeek strongly preferring the first option (Z: 45.8%, $p = 4.15\times10^{-53}$) and avoiding both subsequent options (J: 26.6%, $p = 1.33\times10^{-17}$; Q: 27.6%, $p = 3.29\times10^{-13}$). In contrast, GPT-5 exhibited weaker but still significant order effects ($\chi^2(4, N = 29{,}760) = 16.04$, $p = 0.006$, Cramer's $V = 0.016$), showing a consistent modest preference for option Q across all choice order conditions, reaching significance in the JQZ ($p = 8.28\times10^{-4}$) and ZJQ ($p = 1.04\times10^{-6}$) conditions but not in the QJZ condition ($p = 0.190$), regardless of Q's position in the sequence.

Choice probabilities were also examined on the first trial, where models exhibited near-deterministic behaviour (Fig S10 b). As with the evaluations of choice probability across all trials, only baseline prompting conditions were included. Specifically, DeepSeek showed completely deterministic first-choice behaviour in the QJZ and ZJQ conditions (100% selection of the first-position option in each), while GPT-5 exhibited strong but non-deterministic biases (78.9-82.6% preference rates). DeepSeek demonstrated an extreme first-position bias ($\chi^2(4, N = 744) = 1049.57$, $p = 1.29\times10^{-225}$, Cramer's $V = 0.84$), where in the JQZ condition, 83.5% selected J ($p = 3.27\times10^{-132}$), with only 16.5% choosing Q and 0% choosing Z. In the QJZ and ZJQ conditions, choice was completely deterministic, where 100% of DeepSeek participants (90/90 in each condition) selected the first-position option. On the other hand, GPT-5 showed a qualitatively different pattern, a content-specific preference for option Q that was largely position-independent ($\chi^2(4, N = 744) = 270.00$, $p = 6.37\times10^{-57}$, Cramer's $V = 0.43$). In the JQZ condition, 82.6% chose Q on trial 1 ($p = 1.47\times10^{-127}$). In the ZJQ condition, 78.9% chose Q despite it appearing in second position ($p = 1.79\times10^{-18}$). However, when Q appeared first (QJZ condition), only 7.8% selected it, with 82.2% choosing Z instead.

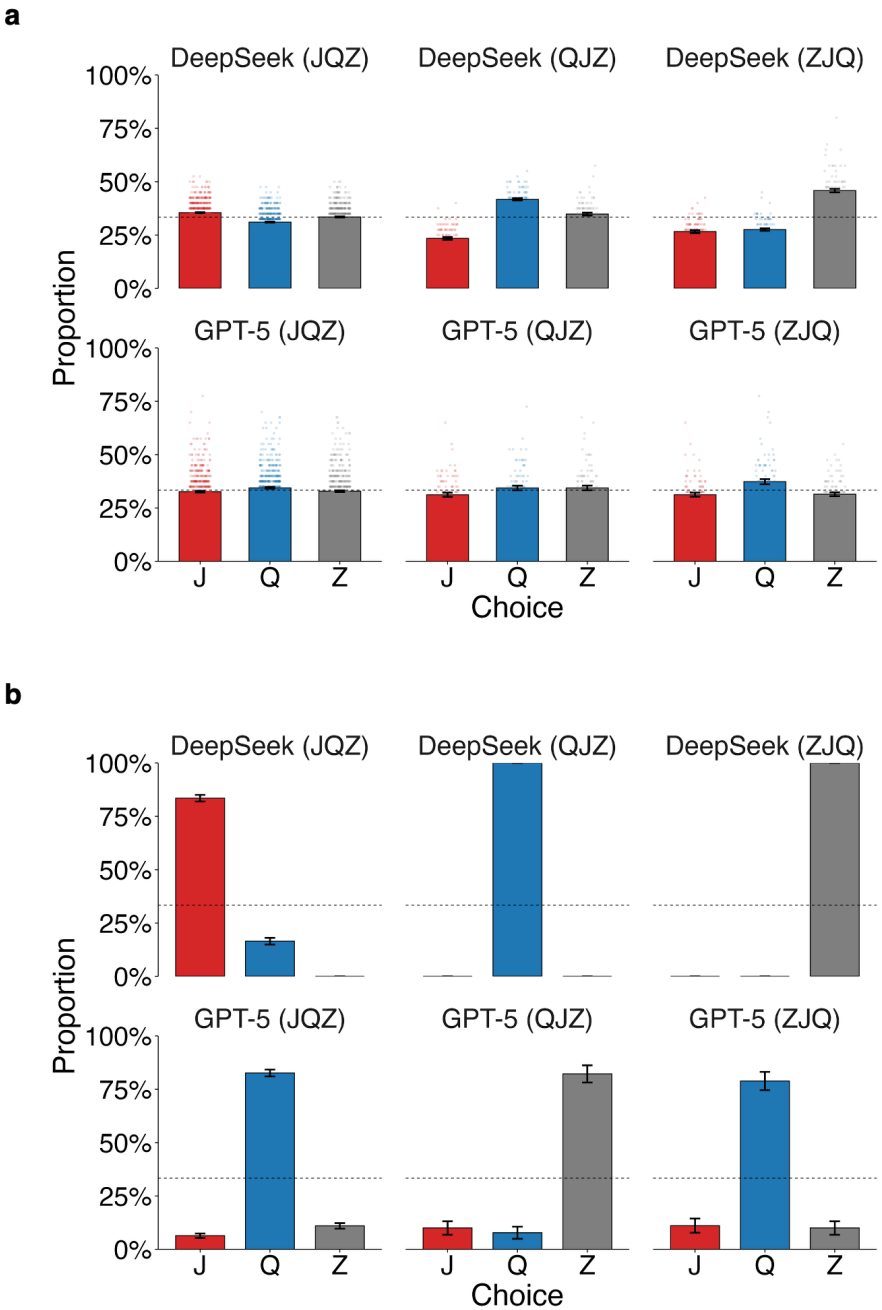


Figure S10: **Effects of order and language on choice proportions by Large Language Models in the rock-paper-scissors game.** **(a)** The effect of recoding choice options on overall choice probability and **(b)** choice probability on the very first trial.

Appendix

Inspection game

Parameter–behaviour regression

Table S3: **Parameter–behaviour regression coefficients for the inspection game.** Standardised regression coefficients b (standard error) from within-group regressions of each behavioural summary on the fitted first-order learning rate (η) and second-order weight (κ). Predictors are log-transformed and then z-scored across the corresponding set. **Bold** indicates $p < 0.05$. SD_η / SD_κ give each group's between-subject standard deviation of the predictors in pooled-z units. DeepSeek-SCoT coefficients are marked † given the unique set.

Group (n)	SD_η / SD_κ	Predictor	κ / p(Switch)	η / opponent tracking	κ / mean payoff
Human (younger) (67)	0.74 / 0.73	η	−0.009 (0.013)	**0.163 (0.024)**	**−0.023 (0.007)**
		κ	**0.101 (0.013)**	**−0.073 (0.024)**	**0.027 (0.007)**
DeepSeek-SCoT† (50)	**1.00 / -**	**η (fictitious)**	**0.018 (0.004)**	**0.057 (0.004)**	0.006 (0.004)
Gemini-SCoT (50)	0.12 / 0.29	η	0.069 (0.038)	0.106 (0.072)	0.042 (0.033)
		κ	0.031 (0.016)	0.037 (0.030)	0.004 (0.013)
GPT-4.1-SCoT (50)	0.07 / 0.03	η	0.017 (0.054)	**0.622 (0.153)**	−0.052 (0.076)
		κ	**0.677 (0.133)**	−0.386 (0.379)	−0.135 (0.187)
GPT-5 (100)	0.05 / 0.17	η	**−0.249 (0.052)**	**1.168 (0.088)**	**−0.207 (0.062)**
		κ	**0.139 (0.016)**	**0.232 (0.028)**	−0.004 (0.019)

Generative posterior predictive check

Table S4: **Generative posterior predictive check for the inspection game.** Each cell is the percentage of the posterior-predictive discrepancy (generated − observed group mean) that falls within the ROPE (±0.05), for the group's winning model run forward from its own simulated history, across the three reported summary statistics (probability of working, switch rate, mean payoff). **Bold** marks statistics accepted as practically reproduced (>97.5% of the discrepancy posterior inside the ROPE).

Group	Winning model	p(Work) (%)	p(Switch) (%)	Payoff (%)
Human (younger)	Influence play	**100**	**100**	**100**
DeepSeek	RL	0	5	0
DeepSeek-SCoT	Fictitious play	17	**100**	**100**
Gemini-SCoT	Influence play	**100**	0	**100**
GPT-4.1	RL	0	64	0
GPT-4.1-SCoT	Influence play	**100**	**100**	**100**
GPT-5	Influence play	**100**	0	**100**

Model comparison and parameter estimates

Full per-group model-comparison and parameter-estimate tables for the inspection game are given below.

Table S5: Model fits for human and LLM groups in the inspection game. Values shown are LOOIC (Leave-One-Out Information Criterion) with standard error in brackets, from the JAGS/Gibbs pipeline. The three learning accounts (reinforcement learning, fictitious play, influence play) were fitted under the [0,1] parameterisation and compared against a memoryless mixed-equilibrium reference. Lower values indicate better out-of-sample predictive accuracy. Winning models for each group are shown in bold. For Gemini (normal prompting) all three learning models collapsed to chance under the [0,1] parameterisation, so the memoryless reference provided the best fit (i.e. no learning winner).

Group	Model	LOOIC (SE)
Human	RL	12385.97 (67.80)
	Fictitious play	12578.76 (62.65)
	Influence play	**12236.42 (71.00)**
	Mixed-equilibrium	12356.39 (68.13)
DeepSeek	**RL**	**707.73 (27.24)**
	Fictitious play	10299.86 (0.36)
	Influence play	10305.45 (0.30)
	Mixed-equilibrium	906.77 (61.06)
DeepSeek-SCoT	RL	10018.38 (32.83)
	Fictitious play	**5387.27 (101.90)**
	Influence play	5400.02 (102.17)
	Mixed-equilibrium	10000.97 (36.11)
Gemini	RL	10065.09 (0.87)
	Fictitious play	10098.79 (1.35)
	Influence play	10084.72 (0.65)
	Mixed-equilibrium	**8121.40 (82.22)**
Gemini-SCoT	RL	7110.03 (92.40)
	Fictitious play	5720.26 (97.99)
	Influence play	**5589.00 (97.07)**
	Mixed-equilibrium	8127.94 (86.85)
GPT-4.1	**RL**	**2556.71 (93.04)**
	Fictitious play	10293.57 (0.40)
	Influence play	10295.48 (0.32)
	Mixed-equilibrium	2852.14 (103.68)
GPT-4.1-SCoT	RL	9956.15 (35.72)
	Fictitious play	10091.26 (27.69)
	Influence play	**8828.51 (70.74)**
	Mixed-equilibrium	10270.74 (12.31)
GPT-5	RL	17721.08 (98.24)
	Fictitious play	16867.38 (108.29)
	Influence play	**16782.49 (109.98)**
	Mixed-equilibrium	19087.71 (75.65)

Table S6: Parameter estimates for human and LLM groups in the inspection game. Values shown are the group-level posterior means for the winning model in each group (η = first-order learning rate; κ = second-order weight; β = inverse temperature). For the human group, the population standard deviation is given in brackets; for the LLM groups the hierarchically estimated between-subject variance was near zero, so only the group-level point estimate is shown. A dash indicates a parameter not present in that group's winning model. Gemini (normal) was best fit by the memoryless mixed-equilibrium reference, whose only parameter is the mixing rate p(Work) = 0.751.

Group	**Winning model**	η	κ	β
Human	Influence play	0.0056 (0.018)	0.0275 (0.050)	0.2954 (0.204)
DeepSeek	RL	0.1351	—	0.9940
DeepSeek-SCoT	Fictitious play	0.5573	—	0.1280
Gemini	Mixed-equilibrium	—	—	—
Gemini-SCoT	Influence play	0.0535	0.0022	0.9261
GPT-4.1	RL	0.0957	—	0.9819
GPT-4.1-SCoT	Influence play	0.1319	0.4416	0.0737
GPT-5	Influence play	0.1163	0.0240	0.1882

Rock-paper-scissors

Model comparison and parameter estimates

Table S7: Model comparison results for the Rock-Paper-Scissors game. Values shown are mean ± standard deviation.

Dataset	**Model**	**negLL**	**AIC**	**BIC**	**PXP**
DeepSeek	**CHASE**	**100.02 ± 7.18**	**210.03 ± 14.37**	**223.97 ± 14.37**	**1**
	Fictitious play	115.09 ± 7.01	234.18 ± 14.02	239.76 ± 14.02	0
	ToMk	114.30 ± 6.65	234.60 ± 13.30	242.96 ± 13.30	0
	Full EWA	114.96 ± 7.00	237.93 ± 14.01	249.08 ± 14.01	0
	Reward Learner	115.10 ± 5.02	234.20 ± 10.04	239.78 ± 10.04	0
	Self-tuning EWA	125.93 ± 3.53	253.86 ± 7.06	256.65 ± 7.06	0
DeepSeek-SCoT	**Fictitious play**	89.98 ± 9.98	**183.96 ± 19.96**	**189.53 ± 19.96**	**0.999**
	CHASE	**87.48 ± 8.94**	184.97 ± 17.87	198.91 ± 17.87	0.001
	ToMk	89.73 ± 10.34	185.47 ± 20.67	193.83 ± 20.67	0
	Reward learner	102.38 ± 8.05	208.75 ± 16.10	214.32 ± 16.10	0
	Self-tuning EWA	114.86 ± 7.32	231.72 ± 14.64	234.51 ± 14.64	0
	Full EWA	114.75 ± 6.75	237.51 ± 13.49	248.66 ± 13.49	0
GPT-5	**ToMk**	104.15 ± 6.51	**214.30 ± 13.02**	**222.67 ± 13.02**	**1**
	CHASE	**103.89 ± 6.07**	217.78 ± 12.14	231.72 ± 12.14	0
	Fictitious play	120.05 ± 5.26	244.11 ± 10.51	249.68 ± 10.51	0
	Reward learner	126.71 ± 3.76	257.43 ± 7.53	263.00 ± 7.53	0
	Full EWA	126.60 ± 2.98	261.20 ± 5.96	272.35 ± 5.96	0
	Self-tuning EWA	132.04 ± 0.14	266.08 ± 0.28	268.86 ± 0.28	0
Human	**CHASE**	**118.77 ± 9.76**	**247.53 ± 19.52**	**261.39 ± 19.57**	**1**
	ToMk	123.34 ± 6.94	252.68 ± 13.88	261.00 ± 13.93	0
	Fictitious play	126.68 ± 6.12	257.35 ± 12.24	262.90 ± 12.27	0
	Reward learner	126.79 ± 6.19	257.57 ± 12.38	263.12 ± 12.41	0
	Self-tuning EWA	128.01 ± 5.36	258.02 ± 10.72	260.79 ± 10.74	0
	Full EWA	126.56 ± 6.63	261.12 ± 13.25	272.21 ± 13.31	0

Table S8: Parameter estimates for human and LLM groups in the RPS task. Values shown are for the winning model in each group. Values are mean estimates with standard deviation in brackets.

Group	α	β	κ	λ	γ
DeepSeek	0.573 (0.273)	3.307 (0.714)	1.755 (0.467)	0.089 (0.696)	0.037 (0.107)
DeepSeek-SCoT	0.217 (0.054)	3.640 (0.430)	—	—	—
GPT-5	0.513 (0.125)	1.420 (0.406)	1.559 (0.539)	—	—
Human	0.542 (0.339)	2.474 (2.359)	1.989 (1.083)	2.109 (1.998)	1.417 (1.910)

Table S9: Effects of order and language for decisions made by Large Language Models in the inspection game. Each subject played a single session of 150 trials. Mean values are presented with standard deviation in brackets where appropriate.

model	**option**	**n**	**payoff**	**p(work)**	**p(work) T1**
DeepSeek	work	50	18.0 (1.4)	0.019	0.680
DeepSeek	x	25	19.2 (1.2)	0.066	1.000
DeepSeek	j	25	18.3 (1.4)	0.026	0.000
Gemini	work	49	14.2 (1.5)	0.751	1.000
Gemini	x	25	10.4 (0.8)	0.859	1.000
Gemini	j	27	18.9 (1.3)	0.012	0.000
GPT-4.1	work	50	18.5 (1.0)	0.055	1.000
GPT-4.1	x	25	18.5 (1.5)	0.006	0.000
GPT-4.1	j	25	18.1 (1.4)	0.002	0.000
GPT-5	work	100	26.6 (1.7)	0.343	0.570
GPT-5	x	25	26.0 (1.7)	0.279	0.080
GPT-5	j	24	25.8 (1.7)	0.271	0.042

Table S10: Effects of order and language for decisions made by Large Language Models in the rock-paper-scissors game. Each subject played a single session of 40 trials. Mean values are presented with standard deviation in brackets where appropriate.

model	**order**	**n**	**payoff**	**p(J)**	**p(Q)**	**p(Z)**	**p(J) T1**	**p(Q) T1**	**p(Z) T1**
DeepSeek	JQZ	564	-0.061 (0.249)	0.355	0.310	0.335	0.835	0.165	0.000
DeepSeek	QJZ	90	-0.049 (0.277)	0.234	0.417	0.348	0.000	1.000	0.000
DeepSeek	ZJQ	90	-0.048 (0.252)	0.266	0.276	0.458	0.000	0.000	1.000
GPT-5	JQZ	564	0.222 (0.116)	0.327	0.344	0.329	0.064	0.826	0.110
GPT-5	QJZ	90	0.222 (0.129)	0.312	0.344	0.344	0.100	0.078	0.822
GPT-5	ZJQ	90	0.238 (0.123)	0.312	0.373	0.314	0.111	0.789	0.100

References

[1] Riemer, M. *et al.* Position: Theory of Mind Benchmarks are Broken for Large Language Models (2025). 2412.19726.

[2] Vidler, A. & Walsh, T. Playing games with Large language models: Randomness and strategy (2025). 2503.02582.

[3] Fontana, N., Pierri, F. & Aiello, L. M. Nicer than Humans: How Do Large Language Models Behave in the Prisoner's Dilemma? *Proceedings of the International AAAI Conference on Web and Social Media* **19**, 522–535 (2025). https://doi.org/10.1609/icwsm.v19i1.35829.