Observer, Not Player: Simulating Theory of Mind in LLMs through Game Observation

Jerry Wang, Ting Yiu Liu
Department of Management Information Systems,
National ChengChi University
111306078@nccu.edu.tw, 113356048@nccu.edu.tw

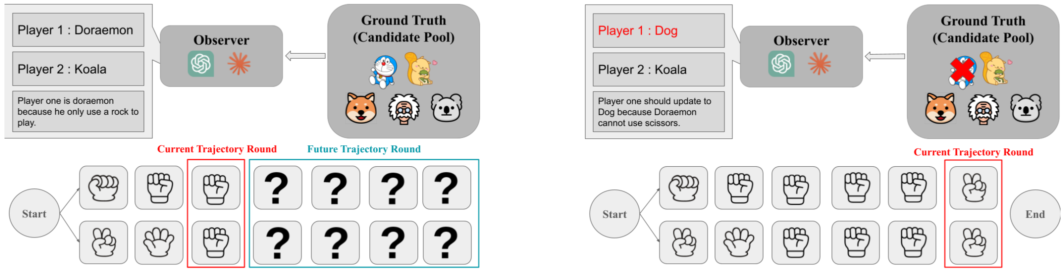
Abstract

We present an interactive framework for evaluating whether large language models (LLMs) exhibit genuine "understanding" in a simple yet strategic environment. As a running example, we focus on Rock–Paper–Scissors (RPS), which, despite its apparent simplicity, requires sequential reasoning, adaptation, and strategy recognition. Our system positions the LLM as an *Observer* whose task is to identify which strategies are being played and to articulate the reasoning behind this judgment. The purpose is not to test knowledge of Rock–Paper–Scissors itself, but to probe whether the model can exhibit mind-like reasoning about sequential behavior. To support systematic evaluation, we provide a benchmark consisting of both static strategies and lightweight dynamic strategies specified by well-prompted rules. We quantify alignment between the Observer's predictions and the ground-truth distributions induced by actual strategy pairs using three complementary signals: Cross-Entropy, Brier score, and Expected Value (EV) discrepancy. These metrics are further integrated into a unified score, the *Union Loss*, which balances calibration, sensitivity, and payoff alignment. Together with a Strategy Identification Rate (SIR) metric, our framework captures not only predictive accuracy but also whether the model can stably identify the latent strategies in play. The demo emphasizes **interactivity, transparency, and reproducibility**. Users can adjust LLM distributions in real time, visualize losses as they evolve, and directly inspect reasoning snippets to identify where and why failures occur. In doing so, our system provides a practical and interpretable proxy for mind-like inference in sequential games, offering insights into both the strengths and limitations of current LLM reasoning.

1 Motivation and Goal

Evaluating what a large language model (LLM) has truly learned in an interactive setting is challenging, because real-world mechanisms are complex and seldom come with explicit explanations. Games provide a useful test bed because their rules are transparent and can be regarded as ground truth. Also, the environment is controllable, and interactions can be reproduced through prompting. For these reasons, games have long served as benchmarks for studying AI behavior.

Recent studies show that LLMs can play games, craft strategies, and adapt to varied environments, spanning board games, repeated social dilemmas, simulations, and program syntheses Bateni and Whitehead [2024], McAleese et al. [2024], Jun et al. [2024], Liu et al. [2024], Wang et al. [2024], Prystawski et al. [2024], Lei et al. [2024]. In most of these works, models act as agents that interact with opponents. Because evaluation often relies only on overall win rate, it is hard to tell whether a model understands the game or merely adapts to observed losses. Win rate compresses pattern recognition, planning, opponent modeling, and luck into a single number, so a high score does not guarantee that the model captures the underlying outcome distribution against each opponent strategy. When the opponent or the initial conditions change, metrics based on win rate become even



(a) Early-round guessess by the LLM Observer

(b) Late-round update of previous strategy

Figure 1: Evolution of the LLM Observer's probabilistic guessess over the course of the match.

less reliable. This issue connects to the broader debate on whether LLMs exhibit *Theory of Mind (ToM)*–like reasoning or simply exploit dataset artifacts and surface cues Kosinski [2024], Ullman [2023].

To obtain a clearer picture of the model's capabilities, we let the LLM act solely as an *Observer*. For each matchup, the model is tasked with predicting the probability distribution over candidate strategies in *Rock–Paper–Scissors (RPS)*, and we compare its guesses against the ground-truth distribution generated by the game engine. To provide the LLM with sufficient information for reasoning, our framework encapsulates the ground-truth rationale, current trajectory round, and trajectory history within a comprehensive *Chain-of-Thought (CoT)* prompt. This unified context leverages recent advances in prompting techniques, which have significantly enhanced LLM reasoning abilities. Specifically, CoT prompting Wei et al. [2022] introduces intermediate reasoning steps into few-shot exemplars, enabling LLMs to solve complex tasks with greater accuracy. Recent paradigms such as Coconut Hao et al. [2024] extend this by supporting reasoning in continuous latent spaces, while Zero-Shot CoT Kojima et al. [2022] demonstrates that even minimal prompt engineering can elicit strong step-by-step reasoning in large models. For practical deployment and real-time interpretability, we adopt standard zero-shot CoT prompting. This allows each round's prediction and reasoning process to be transparently displayed to users, rather than relying on more complex or opaque paradigms such as Coconut.

For evaluation, we operationalize "game understanding" as distributional alignment between the model's guesses and the ground-truth outcome distribution.We quantify this alignment using three proper scoring metrics: *Cross-Entropy (CE)* Kullback and Leibler [1951], Shannon [1948], Goodfellow et al. [2016], the Brier score Brier [1950], and an *Expected-Value discrepancy (EV)* Gneiting and Raftery [2007], Dawid [1984]. To ensure comparability, we apply fixed-bound normalization to the EV term, no further adjustment to the Brier score since it already lies in $[0, 1]$, and grid-wise min–max normalization to the CE term, and then average the three metrics to define the **Union Loss**. A model that genuinely understands the matchup should achieve low values on every component, resulting in a low Union Loss.

We selected Rock–Paper–Scissors for its combination of simplicity and strategic depth. Its transparent rules and non-transitive payoff structure allow us to populate our strategy pool with a diverse set of opponents including static strategies with pure and biased to represent dynamic uniform strategies agents without a fixed plan but sampling with bias, and human-inspired policies such as "win last," "lose last," and "copy last". Moreover, any two strategies admit a closed-form outcome distribution and can be analyzed efficiently via trajectory approximations. The action space also makes RPS ideal for supporting on-laptop demos aligned with a demo-track's goal.

Figure 1a and Figure 1b illustrate how LLMs serve as observers in the gaming environment. In Figure 1a, at the start of a match, the observer issues an initial guess after several rounds. This guess may assign different thought from possibility from both players in this example or even gives strange thoughts in other cases, reflecting a biased hypothesis based on sparse evidence on current trajectory rounds. As the game unfolds (Figure 1b), guesses are continuously compared with the true outcome distribution. By the late rounds, both player provides more gaming history so the previous hypotheses are discarded. Through this dynamic process, we observe how the LLM updates its beliefs and converges on the most likely outcome. This mechanism mirrors how humans form early assumptions and revise them when new information arrives.

Goal of the demo. Our goal is not to chase state-of-the-art benchmarks but to deliver an interactive, transparent and reproducible platform for probing distributional understanding in a simple strategic environment. We provide a real-time web interface that lets users configure static and dynamic opponents, then view the Observer's probability guesses alongside proper-scoring-rule losses. Live visual analytics highlight where and why the model's beliefs diverge from the true outcome distribution. The lightweight implementation promotes community reuse and reproducibility. The design also supports ablation studies—such as varying prompts or temperature settings—and can be extended with methods like change-point detection for strategy switches Adams and MacKay [2007]. Ultimately, the demo demonstrates how an LLM forms, updates and refines its beliefs when interacting with a game-like environment, which also have potential to become benchmarks for theory of mind research.

2 Related Work

2.1 LLM ToM Evaluation

A growing body of work probes whether large language models exhibit theory-of-mind-like competencies Strachan et al. [2024]. Beyond single-task probes, ToMBench systematizes ToM evaluation into 8 task families and 31 abilities to mitigate leakage and subjective grading Chen et al. [2024]. In addition, recent work benchmarks 11 models against children 7–10 years old on richer ToM inventories van Duijn et al. [2023]. Recent position papers argue that many benchmarks capture *literal* ToM (predicting others) rather than *functional* ToM (adapting to new partners), urging interactive settings that test adaptation, calibration, and belief updating Riemer et al. [2025]. Our setup complements this line by treating "understanding" as alignment to *outcome distributions* rather than pass/fail answers, and by making uncertainty explicit through proper scoring rules.

2.2 Interactive Visualization Platforms for Games and Agents

Interactive, code-light tools have long supported transparent debugging and diagnosis. TensorBoard popularized real-time tracking and embedding projections for deep models Abadi et al. [2016], while the What-If Tool introduced point-and-click counterfactuals and per-example analysis without code changes Wexler et al. [2019]. For RL and games, OpenAI Gym standardized interfaces and built-in monitoring/video capture Brockman et al. [2016]; PettingZoo unified multi-agent environments under an AEC API Terry et al. [2021]; OpenSpiel offered a research platform for many games with evaluation utilities Lanctot et al. [2020]; RLCard added card-game suites and visual tools Zha et al. [2020]; Unity ML-Agents provided a general 3D simulation stack with built-in viewers Juliani et al. [2020]. Visualization systems tailored to RL (e.g., Vizarel/DRLViz) further target policy rollouts, memory inspection, and failure analysis Deshpande et al. [2020]. Our dashboard extends this tradition to *distributional* diagnostics in RPS.

2.3 Game-Based Benchmarks for Agent Evaluation

Simple strategic games offer controllable yet revealing tests of planning, opponent modeling, and adaptation. *SmartPlay* curates multiple games (incl. RPS) to isolate nine agent capabilities and multi-turn generalization Wu et al. [2024]. *GameBench* targets LLM strategic reasoning across nine game environments Costarelli et al. [2024], and *AgentBench* evaluates LLMs-as-agents across diverse interactive tasks Liu et al. [2023]. Cooperative and imperfect-information games provide incisive stress tests: the Hanabi Challenge foregrounds belief inference and partner modeling Bard et al. [2020]; Overcooked-AI probes human–AI coordination under time pressure Carroll et al. [2020]. Text-based and gridworld platforms add language and compositionality: TextWorld for language-conditioned RL in generated games Côté et al. [2019], BabyAI for grounded-language curricula Chevalier-Boisvert et al. [2019], and the NetHack Learning Environment for long-horizon, procedural play with rich state/action spaces Küttler et al. [2020]. Recent LLM-focused game suites further emphasize turn-by-turn logging and leaderboarded play on grid-based games such as Tic-Tac-Toe, Connect Four, and Gomoku Topsakal et al. [2024]. Relative to these, our demo centers not on win rate, but on *probabilistic guessing* of outcomes against static/dynamic opponents, exposing where models' beliefs diverge from ground truth.

3 System Overview

Workflow and Steady–State Solver. Figure 2 shows the real-time pipeline. The **Candidate Pool** (App. A) is organized into three distinct classes to capture different behavioral patterns. First, we include *Human (reactive) policies* (X, Y, Z) that update as a function of the opponent's distribution or the previous outcome—namely "win-last," "lose-last," and "copy-last"—to model simple, memory-dependent behaviors. Next, we consider *static strategies* (A, B, C), which are pure Rock, Paper, or Scissors distributions serving as analytically tractable baselines. Finally, the remaining entries (D–P) are *biased dynamic mixtures* with fixed probability vectors that interpolate or bias the pure actions. Only the adaptive strategies require an update map $g_k : \Delta^3 \to \Delta^3$.

For the steady-state approximation, let $s_i^{(t)} \in \Delta^3$ denote player i's mixed strategy at solver iteration t, initialized at $s_i^{(0)} = (1/3, 1/3, 1/3)$. If player i is adaptive (X, Y, Z), we apply a damped fixed-point iteration

$$s_i^{(t+1)} \;=\; \alpha\, g_{k_i}\big(s_{-i}^{(t)}\big) \;+\; (1-\alpha)\, s_i^{(t)}, \qquad \alpha \in (0,1), \tag{1}$$

while static or biased mixture strategies keep $s_i^{(t+1)} = s_i^{(t)}$.

In the fully co-adaptive regime, we simulate two adaptive agents through the coupled updates

$$s_1^{(t+1)} = \alpha\, g_{k_1}\big(s_2^{(t)}\big) \;+\; (1-\alpha)\, s_1^{(t)}, \tag{2}$$

$$s_2^{(t+1)} = \alpha\, g_{k_2}\big(s_1^{(t)}\big) \;+\; (1-\alpha)\, s_2^{(t)}. \tag{3}$$

Iterations stop when

$$\|s_i^{(t+1)} - s_i^{(t)}\|_1 < 10^{-4}, \tag{4}$$

for all adaptive players or when a preset cap is reached. We empirically observe convergence for the tested settings (formal guarantees are out of scope). Let (s_1^*, s_2^*) be the resulting steady state; the outcome probabilities $\Pr[\text{win}], \Pr[\text{draw}], \Pr[\text{loss}]$ are then computed under this stationary-mixing approximation and fed to the Observer's loss.

The current game state and cumulative history are provided to the **Prompt Module**, which constructs a four-part chain-of-thought (CoT) prompt consisting of: (1) candidate information, (2) role specification, (3) previous game trajectory results, and (4) a request for the model to predict the next player action with explicit reasoning. The CoT prompt enables the language model to articulate its decision-making process, thereby providing transparency into *why* a particular choice was made. Furthermore, by encapsulating the full game context within the prompt, our approach ensures that evaluation measures the model's reasoning in context, rather than relying on any latent background knowledge. Specific prompt we used for experiment can be found in Appendix B.

The completed prompt is routed to the **Observer**, an LLM selected at run time, executing the task and generating guesses for both players, with all metrics streamed to the **Evaluation Dashboard**, a lightweight web interface that renders updated result for current model loss, enabling real-time prompt ablations or temperature sweeps.

4 Evaluation Mechanisms

To judge whether an LLM *truly understands* the game—rather than merely fitting surface patterns—we evaluate its predictions at both the distributional and payoff levels via several metrics below. For each matchup, let $p^\star = (p^\star_{\text{win}}, p^\star_{\text{draw}}, p^\star_{\text{loss}})$ denote the ground-truth outcome distribution, and let $\hat{p}$ denote the probability distribution over win, draw, and loss outcomes as predicted by the LLM Observer after modeling the interaction between the two players' strategies.

We report three complementary metrics. First, CE measures the information-theoretic "surprise" when the true distribution is encoded by the model's prediction, rewarding guesses that place high probability on the correct outcomes:

$$\mathrm{CE}(p^\star, \hat{p}) = -\sum_{c \in \{\text{win},\text{draw},\text{loss}\}} p^\star_c \, \log(\hat{p}_c + \varepsilon), \tag{5}$$

where the constant ε prevents numerical underflow. Note that CE is not guaranteed to be zero even when the predicted distribution perfectly matches the true distribution: unless the ground truth is

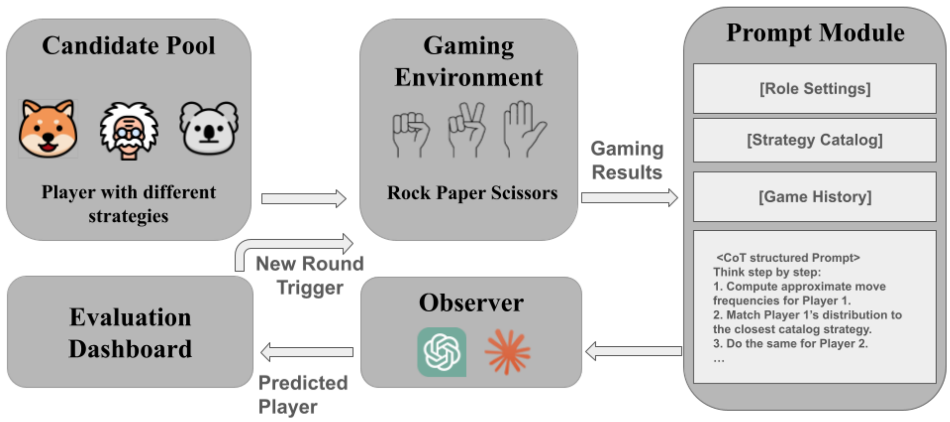


Figure 2: **System pipeline.** Solid arrows indicate the data flow from user-selected strategies in the Candidate Pool, through the RPS engine and Prompt Module, to the LLM Observer and the real-time Evaluation Dashboard.

deterministic (e.g., one outcome with probability 1), CE will equal the entropy of the true distribution, which is strictly positive.

Second, the *Brier score* penalizes both mis-ranking and mis-calibration of probabilities, serving as a calibration-sensitive proper scoring rule:

$$\text{Brier}(p^\star, \hat{p}) = \sum_{c \in \{\text{win,draw,loss}\}} (\hat{p}_c - p^\star_c)^2. \tag{6}$$

Whereas CE and Brier together capture distributional alignment (i.e., whether the model predicts the full outcome distribution correctly and with proper confidence), they may still overlook systematic errors in payoff assessment.

To address this, we further compute the *Expected-Value discrepancy* (EVLoss), which directly checks whether the model over- or underestimates the net advantage implied by its guesses:

$$\text{EV}(p) = \frac{p_{\text{win}} - p_{\text{loss}}}{100},$$
$$\text{EVLoss}(p^\star, \hat{p}) = \left(\text{EV}(p^\star) - \text{EV}(\hat{p})\right)^2. \tag{7}$$

The theoretical range of $\text{EV}(p)$ is $[-1, 1]$, since wins and losses can differ by at most 100 out of 100 matches. Therefore, the discrepancy $\text{EV}(p^\star) - \text{EV}(\hat{p})$ lies in $[-2, 2]$, and $\text{EVLoss} \in [0, 4]$. The lower bound 0 occurs when the predicted and true expected values coincide, while the upper bound 4 occurs when they are maximally opposed (e.g., $\text{EV}(p^\star) = 1$ and $\text{EV}(\hat{p}) = -1$).

Finally, we combine the three components into a single metric, *Union Loss*, by averaging normalized scores:

$$\text{Union}(p^\star, \hat{p}) = \frac{1}{3}\left[\text{CE}_{\text{norm}}(p^\star, \hat{p}) + \text{Brier}_{\text{norm}}(p^\star, \hat{p}) + \text{EVLoss}_{\text{norm}}(p^\star, \hat{p})\right]. \tag{8}$$

For comparability, we normalize each component differently according to its natural range: (i) EVLoss is divided by 4.0, its theoretical maximum; (ii) Brier already lies in $[0, 1]$ and is left unchanged; (iii) CE has no fixed bound and is normalized via grid-wise min–max scaling across all matchups:

$$\text{CE}_{\text{norm}}(x) = \begin{cases} \frac{x - \min(\text{CE})}{\max(\text{CE}) - \min(\text{CE})}, & \max(\text{CE}) \neq \min(\text{CE}), \\ 0.5, & \max(\text{CE}) = \min(\text{CE}). \end{cases} \tag{9}$$

These metrics jointly evaluate distinct aspects of model reasoning, each corresponding to an essential component of theory of mind. Specifically, Cross-Entropy and Brier score probe the *behavioral layer*:

does the model accurately predict the full outcome distribution and express appropriately calibrated confidence? In contrast, EV Loss targets the *utility layer* to show whether the model infers both the direction and magnitude of the expected payoff. A low Union Loss thus indicates that the model aligns with both action prediction (cognition) and utility inference—the dual capacities fundamental to robust theory-of-mind–style game understanding.

To provide better intuition of how the framework operates, we visualize the outcome space as a heatmap. Figure 3 illustrates how the loss behaves when the LLM makes different predictions against the ground-truth strategy pair. Each block shows the corresponding loss relative to the actual outcome, with color intensity encoding the degree of deviation. As an illustrative example, we highlight a combat scenario between strategy A and strategy B, where the heatmap clearly demonstrates that mismatched guesses incur larger penalties, whereas accurate predictions correspond to lower loss. This visualization offers an intuitive way for human readers to assess the quality of the Observer Module's reasoning across the candidate pool and serves as a convenient diagnostic tool.

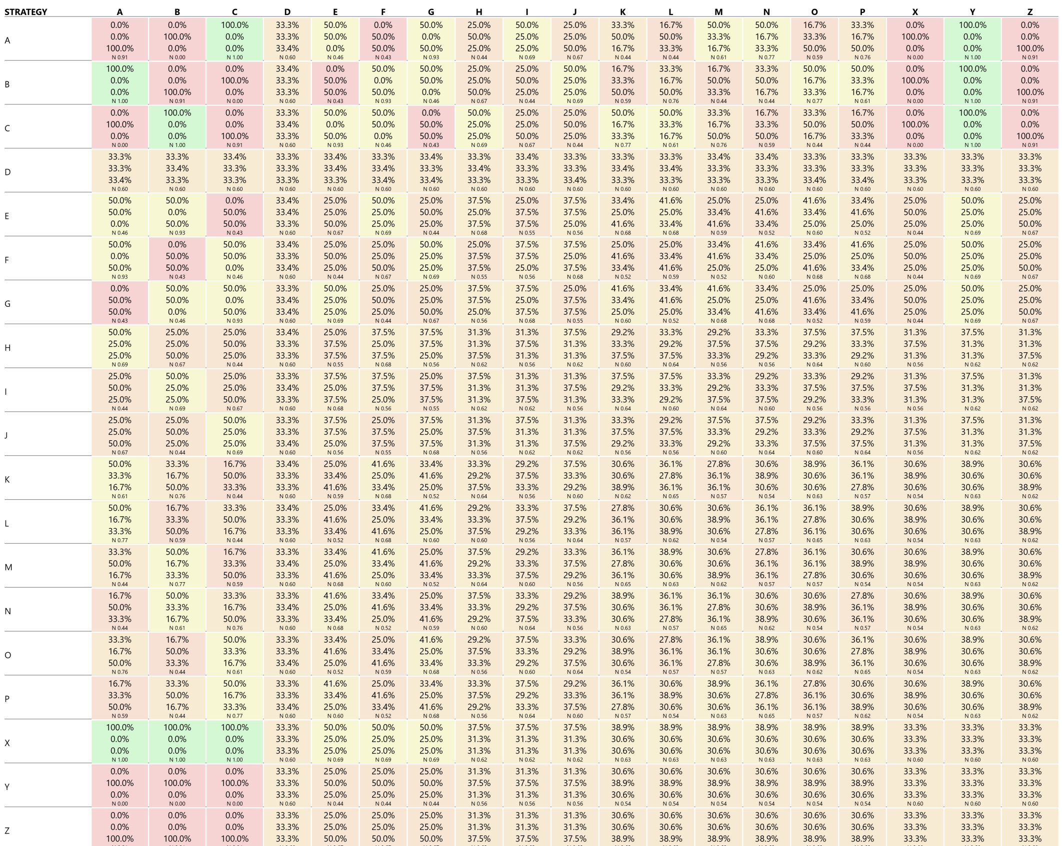


Figure 3: Heatmap visualization of loss values across strategy matchup. Green cells indicate predictions that closely match the actual outcome distribution, yellow to orange denote moderate deviations, and red highlights severe mismatches.

5 Experimental Setting

Our study is guided by two central research questions (RQs). First, *do LM reasoning traces reliably yield correct sequential strategy inferences?* Second, *how do our proposed metrics relate to mind-like inference?* To address these questions, we evaluate each agent under the three interaction patterns introduced in Sec. 3.

Agents. We test three recent instruction-tuned models GPT-4O-MINI, O3, and CLAUDE 3.7 SONNET *(Claude)* OpenAI et al. [2024], Anthropic [2025]. These models were selected to provide coverage across (i) different providers (OpenAI vs. Anthropic), (ii) different cost–latency tiers (lightweight "mini" vs. reasoning-optimized "o3"), and (iii) different design philosophies in alignment and reasoning. This diversity allows us to examine whether distributional understanding is consistent across architectures, scales, and training paradigms, rather than being confined to a single family of models.

Ground truth. We provided Table 3 in Appendix A as the reference set of ground-truth strategies for all player roles. Each role corresponds to a predefined distribution over rock, paper, and scissors moves, ranging from pure strategies to mixed or biased ones, which serve as the canonical answer space against which the LLMs generate their guesses.

LLM prompting. Each model is prompted using a chain-of-thought (CoT) format, as illustrated in Appendix B. The prompt includes the cumulative transcript up to the current round and is issued at a decoding temperature of 0.2 with TOP-P=2.0. For each round, the model outputs its inferred strategies for both players (`guess_s1` and `guess_s2`) along with a confidence score. Every 20 rounds, it additionally produces a reasoning summary composed of 3–5 semicolon-separated phrases to facilitate monitoring of its evolving internal logic. To stabilize initial behavior, the model undergoes a warm-up phase in which it observes 10 prior rounds before making predictions. During inference, only the 50 most recent rounds are available as memory (history limit). The model is free to generate arbitrarily long reasoning but must conclude with the expected structured output.

Match-up configurations and variables. We design three representative match-up regimes to expose models to distinct strategic dynamics:

(1) **Dynamic vs. Static (H vs. C)**: H is a rock-biased player with distribution, while C is a pure paper player with distribution.

(2) **Dynamic vs. Dynamic (N vs. G)**: N favors paper primarily and scissors secondarily, while G mixes paper and scissors.

(3) **Dynamic vs. Psychological (D vs. Y)**: D plays paper-primary, scissors-secondary, whereas Y is adaptive, choosing the counter-move to the opponent's most recent bias.

We treat the choice of *model* as the independent variable, with three levels {GPT-4o-mini, GPT-o3, Claude-3.7}, and the match-up regime as the manipulated condition of true strategies, also with three levels {H vs C, N vs G, D vs Y}, yielding $3 \times 3 = 9$ total experimental conditions. All other parameters are controlled for comparability: the number of rounds is fixed at 200; a warm-up of 10 rounds allows the model to observe before issuing guesses; the history limit is set to 50 as the maximum number of past rounds the model may reference; and a reasoning interval of 20 requires the model to emit a brief reasoning summary every 20 rounds.

Evaluation metrics. For each pair of guesses $(p^\star, \hat{p})$, we evaluate performance using the metrics defined in Sec. 4: (1) Cross-Entropy, (2) Brier score, (3) EV discrepancy, and (4) their aggregate Union Loss. Also, we consider an (5)*explicit-commitment* metric (SIR) that requires correct identification of *both* players' strategies.

6 Demo Scenarios and Formative Findings

6.1 Evaluation of LLM Reasoning Traces for Sequential Strategy Inference (RQ1)

Reasoning traces are often taken as evidence of "understanding," yet it remains unclear whether such narratives *reliably* lead to correct answers in sequential settings. To address this gap, we operationalize reliability through three complementary lenses: loss-based signals, an explicit-commitment metric (SIR), and model-provided reasoning. Together, these perspectives allow us to assess whether reasoning traces consistently yield correct sequential inferences. Agreement across these lenses would indicate convergent validity for our evaluation and, importantly, capture competencies that are necessary but not sufficient for mind-like inference.

Figure 4 and Figure 5 reveal clear round-wise differences and convergence patterns across the three models. Among these models, *o3* exhibits a characteristic profile with a brief initial transient, a rapid decline to a low level, and a stable plateau with only small fluctuations. This trend is evident in the N–G matchup, where the Brier loss approaches zero after roughly 40–60 rounds and remains near

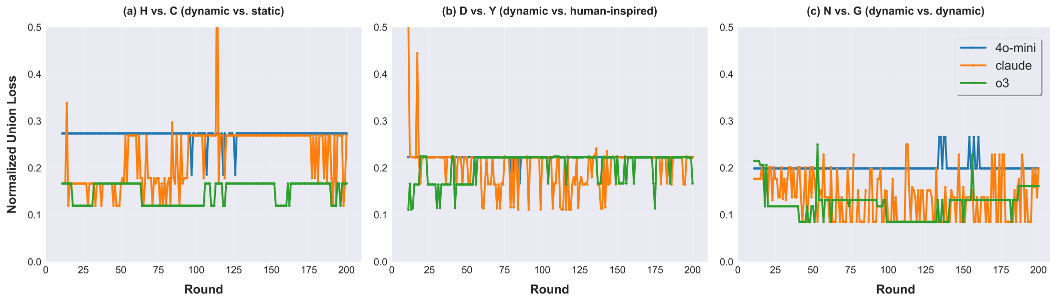


Figure 4: Comparison of Union Loss across models in different settings.

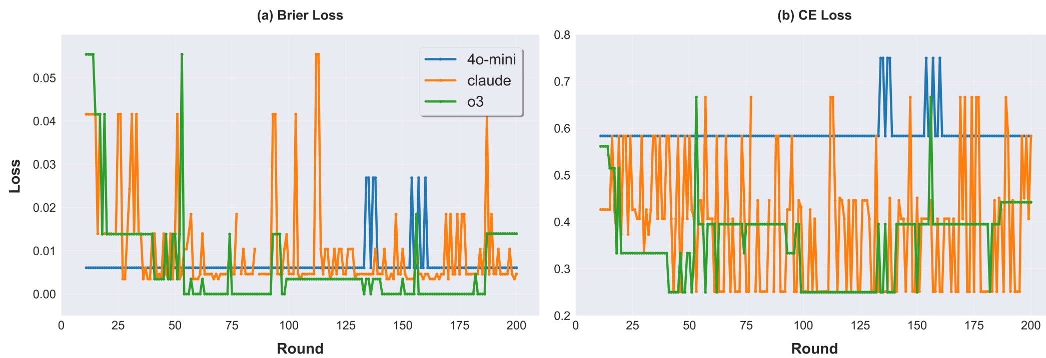


Figure 5: Brier loss and normalized Cross-Entropy loss for the N vs. G matchup

that value, while the normalized Cross-Entropy decreases monotonically between rounds 50–100 and then stays low with only minor excursions. *Claude* shows a mid-range yet high-variance profile. Although its losses tend to decline in the middle-to-late segments, the curves contain frequent spikes of varying magnitude, indicating high sensitivity to local histories and limited stability. In contrast, *GPT-4o-mini* remains almost invariant across rounds. In the N–G matchup, the Brier loss stays as a nearly flat, low-amplitude baseline, whereas the normalized Cross-Entropy remains at a relatively high constant level with a pronounced spike only near rounds 140–160. This combination of a low and steady Brier loss together with a high and flat Cross-Entropy suggests that its predictive distribution is barely updated. The model behaves as if it emits a fixed, smoothed prior, thereby allocating insufficient probability mass to the true class and being penalized under log loss. Taken together, these results provide a direct answer to **RQ1**. We find that only *o3* demonstrates reliable sequential strategy inference, as indicated by rapid convergence and stability across both loss metrics. *Claude* shows partial but unstable reliability due to high variance, while *GPT-4o-mini* fails to adapt its predictions and therefore does not yield reliable inferences.

6.2 Relation of Our Metrics to Mind-like Inference (RQ2)

To assess whether our metrics capture *ToM-relevant* competencies (rather than ToM itself), we triangulate the loss-based trajectories with a textual commitment check—the Strategy Identification Rate (SIR)—and qualitative analysis of the models' own reasoning 2, see Appendix C for full tables.

Across Figure 4, Table 1 (SIR), and representative rationale excerpts (App. C), we observe a consistent model ordering and round-wise dynamics.

o3. o3's rationales follow a systematic pattern: it first quantifies round-wise frequencies, then maps them to the strategy catalog by citing proximity (e.g., smallest distance to *G* for *Paper+Scissors*), and explicitly *rules out* dynamic rules with concrete evidence (e.g., "zero Rock precludes X/Y/Z"). This disciplined style yields *stable static attributions* as the window expands, high SIR in the Static–Dynamic and Dynamic–Dynamic settings (see Table 1), and *low, comparatively smooth* Union Loss in Figure 4(a,b). Even when facing Random vs. Counter-Last (D vs. Y), where success requires tracking *opponent-conditioned, lag-1 contingencies*, o3 remains comparatively strongest, although all models show elevated loss and near-zero SIR (Figure 4c; Table 1).

Table 1: **Strategy Identification Rate (%).** Values denote the percentage of rounds in which the model *explicitly* identified *both* players' true strategies (not a "lowest-loss" share). Each match-up has 200 rounds.

Model	Static vs. Dynamic	Dynamic vs. Dynamic	Human-inspired vs. Dynamic
GPT-4o-mini	0.0%	0.0%	0.0%
o3	57.5%	41.5%	0.5%
Claude 3.7	21.5%	0.0%	1.0%

Claude 3.7. Claude's explanations consistently report aggregate proportions and align them to *static* templates (e.g., "closest to M/N") while often stating "no clear dynamic pattern." This behavior partially reduces loss in Static–Dynamic (Figure 4a) with a correspondingly *moderate* SIR in Table 1, but it degrades in Dynamic–Dynamic (Figure 4b) where sensitivity to opponent-conditioned reactions is required; there SIR collapses to near zero (Table 1). The oscillation between adjacent static labels without engaging the sequential dependency is consistent with its *mid-level, higher-variance* loss curves.

GPT-4o-mini. GPT-4o-mini frequently exhibits *description–attribution inconsistency* and unstable commitments, e.g., labeling a distribution as "Paper-biased" while concluding it "matches Rock-biased (H)," and alternating among Rock-biased, Paper-biased, and dynamic Z. These internally inconsistent or templated narratives rarely culminate in correct dual attributions, producing *near-zero SIR* across settings (Table 1) and *persistently high* Union Loss with limited round-wise improvement (Figure 4). In D vs. Y, it fails to track the one-step, opponent-conditioned countering rule, mirroring the general pattern seen across models.

Table 2: Representative reasoning snippets (abridged) produced by each model for the N vs. G matchup.

Model	Round	Reasoning
o3	40	P2 0% R, ∼50/50 P+S; matches G; no adaptive pattern.
GPT-4o-mini	60	P1 prefers Paper; this aligns with Rock-biased. P2 mirrors (Z).
Claude	140	P1 60% Paper, 28% Scissors; P2 48% Scissors, 44% Paper; no clear dynamic pattern.

Taken together, these findings provide a direct answer to **RQ2**. Our metrics—loss calibration, explicit commitments (SIR), and reasoning consistency—jointly capture competencies that approximate *mind-like inference*. They reflect necessary building blocks such as belief updating, probability calibration, and stable strategy identification, which together offer a practical proxy for Theory of Mind without claiming full equivalence. Additional reasoning tables and excerpts are provided in App. C for completeness.

7 Conclusion and Limitations

In this work, we introduced an interactive and reproducible demo for probing LLM game understanding through Rock–Paper–Scissors, positioning the model as an *Observer* rather than a player. Our framework evaluates distributional alignment between predictions and ground-truth outcomes using a principled Union Loss that integrates Cross-Entropy, Brier score, and Expected-Value discrepancy, thereby capturing both behavioral and utility aspects of reasoning. Real-time visualization and full control over prompts, memory, and candidate strategies make the system transparent, extensible, and suitable as a benchmark for future comparative studies. While our current evaluation is limited to simple environments such as RPS and focuses primarily on surface-level reasoning traces, the framework offers a foundation for broader exploration. Future work could extend it to more complex, multi-agent, or strategic games, and incorporate deeper diagnostics to assess reasoning chains more systematically. We view this work as an initial step toward building interactive, aiming to create interpretable testbeds for understanding the reasoning capabilities of large language models.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning, 2016. URL https://arxiv.org/abs/1605.08695.

Ryan P Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Anthropic. Claude 3.7 sonnet system card. Technical report, Anthropic, 2025. URL https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf.

Nolan Bard, Jakob N. Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H. Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, Iain Dunning, Shibl Mourad, Hugo Larochelle, Marc G. Bellemare, and Michael Bowling. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216, March 2020. ISSN 0004-3702. doi: 10.1016/j.artint.2019.103216. URL http://dx.doi.org/10.1016/j.artint.2019.103216.

Bahar Bateni and Jim Whitehead. Language-driven play: Large language models as game-playing agents in slay the spire. In *Proceedings of the 19th International Conference on the Foundations of Digital Games*, FDG '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400709555. doi: 10.1145/3649921.3650013. URL https://doi.org/10.1145/3649921.3650013.

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016. URL https://arxiv.org/abs/1606.01540.

Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination, 2020. URL https://arxiv.org/abs/1910.05789.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. Tombench: Benchmarking theory of mind in large language models, 2024. URL https://arxiv.org/abs/2402.15052.

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. Babyai: A platform to study the sample efficiency of grounded language learning, 2019. URL https://arxiv.org/abs/1810.08272.

Anthony Costarelli, Mat Allen, Roman Hauksson, Grace Sodunke, Suhas Hariharan, Carlson Cheng, Wenjie Li, Joshua Clymer, and Arjun Yadav. Gamebench: Evaluating strategic reasoning abilities of llm agents, 2024. URL https://arxiv.org/abs/2406.06613.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Ruo Yu Tao, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. Textworld: A learning environment for text-based games, 2019. URL https://arxiv.org/abs/1806.11532.

A. Philip Dawid. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A*, 147(2):278–292, 1984.

Shuby Deshpande, Benjamin Eysenbach, and Jeff Schneider. Interactive visualization for debugging rl, 2020. URL https://arxiv.org/abs/2008.07331.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, 2016.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Arthur Juliani, Vincent-Pierre Berges, Ervin Teng, Andrew Cohen, Jonathan Harper, Chris Elion, Chris Goy, Yuan Gao, Hunter Henry, Marwan Mattar, and Danny Lange. Unity: A general platform for intelligent agents, 2020. URL https://arxiv.org/abs/1809.02627.

He Jun, Ziyan Lin, Jian Zhang, Ziqi Sun, and Lingfeng Xu. Can large language models play games? a case study. *arXiv preprint arXiv:2403.05632*, 2024. URL https://arxiv.org/abs/2403.05632.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213, 2022.

Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45), October 2024. ISSN 1091-6490. doi: 10.1073/pnas.2405460121. URL http://dx.doi.org/10.1073/pnas.2405460121.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Heinrich Küttler, Nantas Nardelli, Alexander H. Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, and Tim Rocktäschel. The nethack learning environment, 2020. URL https://arxiv.org/abs/2006.13760.

Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. Openspiel: A framework for reinforcement learning in games, 2020. URL https://arxiv.org/abs/1908.09453.

Xin Lei, Zhen Liu, Ziqiang Zhang, et al. Instruction-driven game engines on large language models (idge). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://arxiv.org/abs/2404.00276.

Shuyue Liu, Jingyuan Shi, Xiangwen Zhu, et al. lmgame-bench: How good are llms at playing games? *arXiv preprint arXiv:2505.15146*, 2024. URL https://arxiv.org/abs/2505.15146.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating llms as agents, 2023. URL https://arxiv.org/abs/2308.03688.

Natalie McAleese, Niels Goudriaan, Nicholas Gray, Tushar Taneja, and et al. Playing repeated games with large language models. *Nature Human Behaviour*, 8:1–9, 2024. URL https://www.nature.com/articles/s41562-025-02172-y.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Andrzej Prystawski, Kaixuan Yang, Huanru Liu, et al. From code to play: Benchmarking program search for games using large language models. *arXiv preprint arXiv:2412.04057*, 2024. URL https://arxiv.org/abs/2412.04057.

Matthew Riemer, Zahra Ashktorab, Djallel Bouneffouf, Payel Das, Miao Liu, Justin D. Weisz, and Murray Campbell. Position: Theory of mind benchmarks are broken for large language models, 2025. URL https://arxiv.org/abs/2412.19726.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7): 1285–1295, 2024. ISSN 2397-3374. doi: 10.1038/s41562-024-01882-z. URL https://doi.org/10.1038/s41562-024-01882-z.

J. K. Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis Santos, Rodrigo Perez, Caroline Horsch, Clemens Dieffendahl, Niall L. Williams, Yashas Lokesh, and Praveen Ravi. Pettingzoo: Gym for multi-agent reinforcement learning, 2021. URL https://arxiv.org/abs/2009.14471.

Oguzhan Topsakal, Colby Jacob Edell, and Jackson Bailey Harper. Evaluating large language models with grid-based game competitions: An extensible llm benchmark and leaderboard, 2024. URL https://arxiv.org/abs/2407.07796.

Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks, 2023. URL https://arxiv.org/abs/2302.08399.

Max van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit, and Peter van der Putten. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. In Jing Jiang, David Reitter, and Shumin Deng, editors, *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 389–402, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-1.25. URL https://aclanthology.org/2023.conll-1.25/.

Qiyuan Wang, Wen He, Jing Li, et al. Adaptive strategy automation with large language models in paradoxical games. *Physical Review Research*, 7(2):L022012, 2024. URL https://link.aps.org/doi/10.1103/PhysRevResearch.7.L022012.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.

James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viegas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, page 1–1, 2019. ISSN 2160-9306. doi: 10.1109/tvcg.2019.2934619. URL http://dx.doi.org/10.1109/TVCG.2019.2934619.

Yue Wu, Xuan Tang, Tom M. Mitchell, and Yuanzhi Li. Smartplay: A benchmark for llms as intelligent agents, 2024. URL https://arxiv.org/abs/2310.01557.

Daochen Zha, Kwei-Herng Lai, Yuanpu Cao, Songyi Huang, Ruzhe Wei, Junyu Guo, and Xia Hu. Rlcard: A toolkit for reinforcement learning in card games, 2020. URL https://arxiv.org/abs/1910.04376.