GameBench: Evaluating Strategic Reasoning Abilities of LLM Agents

Anthony Costarelli* Mat Allen* Roman Hauksson* Grace Sodunke* Suhas Hariharan Carlson Cheng Wenjie Li Joshua Clymer Arjun Yadav

TL;DR

We introduce GameBench, a cross-domain benchmark evaluating the strategic reasoning ability of large language models (LLMs) as agents by having them compete against each other in a suite of nine varied, text-based, uncommon games.

We test GPT-3.5 and GPT-4 instantiated with and without two scaffolding methods, Chain of Thought (**CoT**) and Reasoning via Planning (**RaP**), along with a human baseline. We find that human play outperforms all configurations followed by GPT-4 scaffolded with Reasoning via Planning.

Games Played and Strategy Types

We identified six orthogonal components of strategic reasoning and selected a suite of nine board, card, and social games that collectively span these dimensions. Due to a low online presence, we believe these games are less represented in LLMs' training corpuses.

- Air, Land, Sea (**ALS**)
- Arctic Scavengers (**AS**)
- Are You the Traitor? (**AYT**)
- Codenames (**CN**)
- Hive (**HV**)
- Pit (**PT**)
- Santorini (**SN**)
- Two Rooms and a Boom (**TRB**)
- Sea Battle (**SB**)

Reasoning Category	Total	Games
Abstract Strategy	6	ALS, AS, CN, HV, SN, SB
Non-Deterministic	3	AS, TRB, SB
Hidden Information	3	AS, AYT, TRB
Language Communication	4	AYT, CN, PT, TRB
Social Deduction	2	AYT, TRB
Cooperation	4	AYT, CN, SB, TRB

Results

We evaluate the following configurations of GPT-3.5 and GPT-4 across our game suite, comparing against both human and random baselines:

- gpt-3
- gpt-3-cot
- gpt-4
- gpt-4-cot
- gpt-4-rap

Our analysis reveals that the human baseline (1.76) significantly outperforms all LLM configurations, while base GPT-4 (-0.89) unexpectedly performs below the random baseline (-0.50).

Figure 1. Overall skill rating for each agent (bootstrapped)

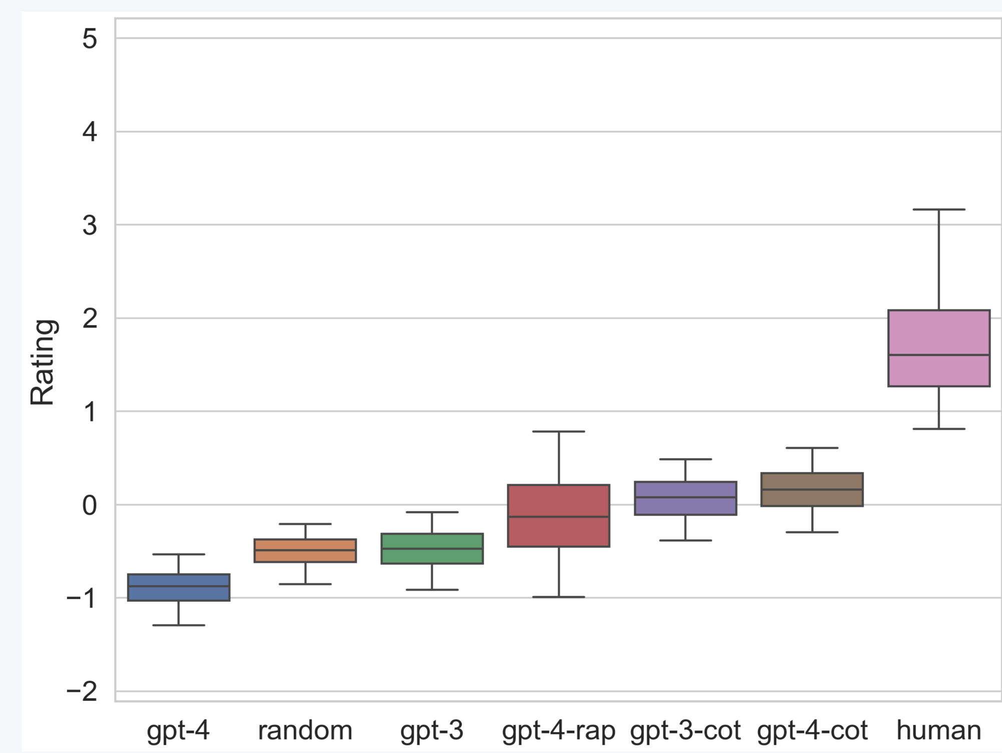
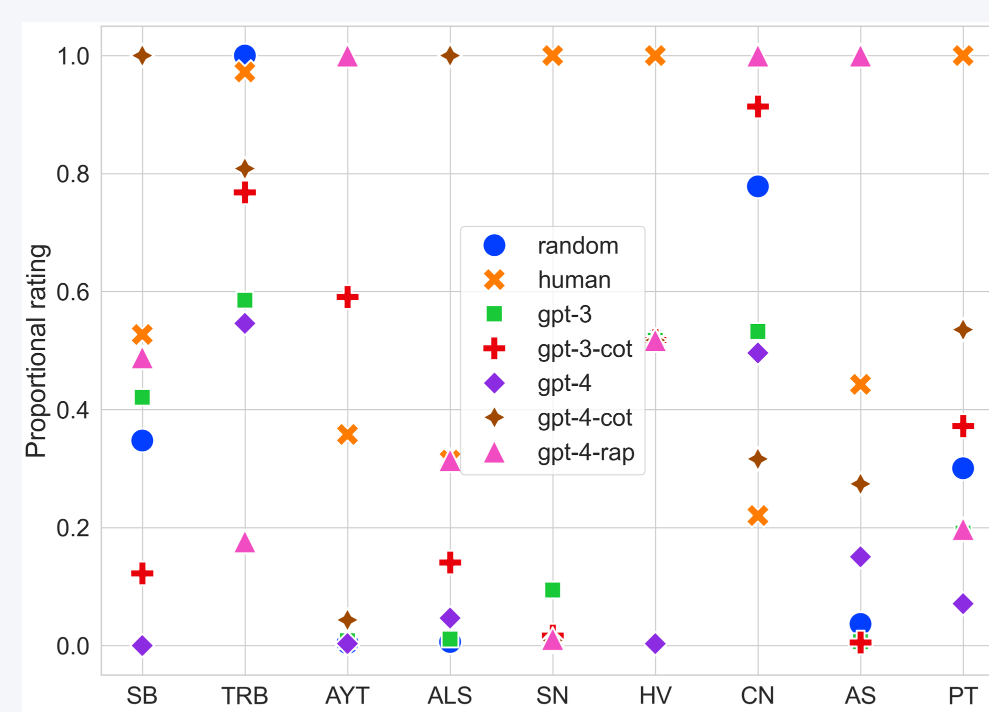


Figure 2. Agent skill ratings per game (as proportion of best rating)



Performance Aggregation

We use the *Bradley–Terry model* to convert match results into overall skill ratings for each agent across each game. Unlike the Elo model, the *Bradley-Terry model* assumes skill level does not change over time, matching the frozen capabilities of a given agent. In addition, it also enables the comparison of agent that never competed with each other.

- Each agent is assigned a rating parameter β that represents their skill level
- For any two agents i and j, probability of i winning against j is modeled as

$$P(i > j) = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}}$$

- To handle varying numbers of matches between games, matches are weighted inversely to number of matches per game: $w_i = \frac{1}{N_X}$ for match i in game X
- Bootstrapping with 10,000 samples provides confidence intervals, selecting matches proportional to their weights
- Final skill ratings ratings are the means of the bootstrapped parameter distributions

Key Contributions

- Created GameBench, a novel framework evaluating strategic reasoning across multiple domains using deliberately out-of-distribution games
- Evaluated state-of-the-art LLMs (GPT-3.5, GPT-4) and scaffolding techniques (Chain-of-Thought, Reasoning Via Planning) against human and random baselines
- Demonstrated that while scaffolding methods improve performance, even enhanced LLMs fall short of human-level strategic reasoning

Future Work

- Design entirely novel games to ensure full out-of-distribution testing
- Test additional LLMs and scaffolding methods
- Analyze and test for the subskills involved in strategic reasoning to identify bottlenecks for superhuman performance in strategic domains

https://gamebench-website.vercel.app/

Language Gamification Workshop @ NeurIPS 2024

acostarelli@olin.edu, matdallen@gmail.com, romanhauksson@gmail.com