

EBR-bench update

2026-10-07

Contents

Overview	1
GPT-6 Astra scores perfectly on EBR-bench by exploiting a powerful combination of cards	1
Multi-agent scaffolds might help models explore, but not improve	8
EBR-bench moving forward	8
About the authors	12
Related work	13
Related topics	15

Overview

EBR-bench is a benchmark that we designed to test models’ ability to learn from experience by playing the board game Earthborne Rangers. The game is complex, with a single playthrough taking human players approximately 2-4 hours to complete, making it useful for assessing a model’s ability to complete complex multi-step tasks, referred to as long-horizon tasks.

With EBR-bench, we **previously found** that models often struggled to match the **human ability** to flexibly explore different decks of cards, optimize tactical play over time, and improve their overall score with experience. We’ve continued to experiment with it and present two new findings.

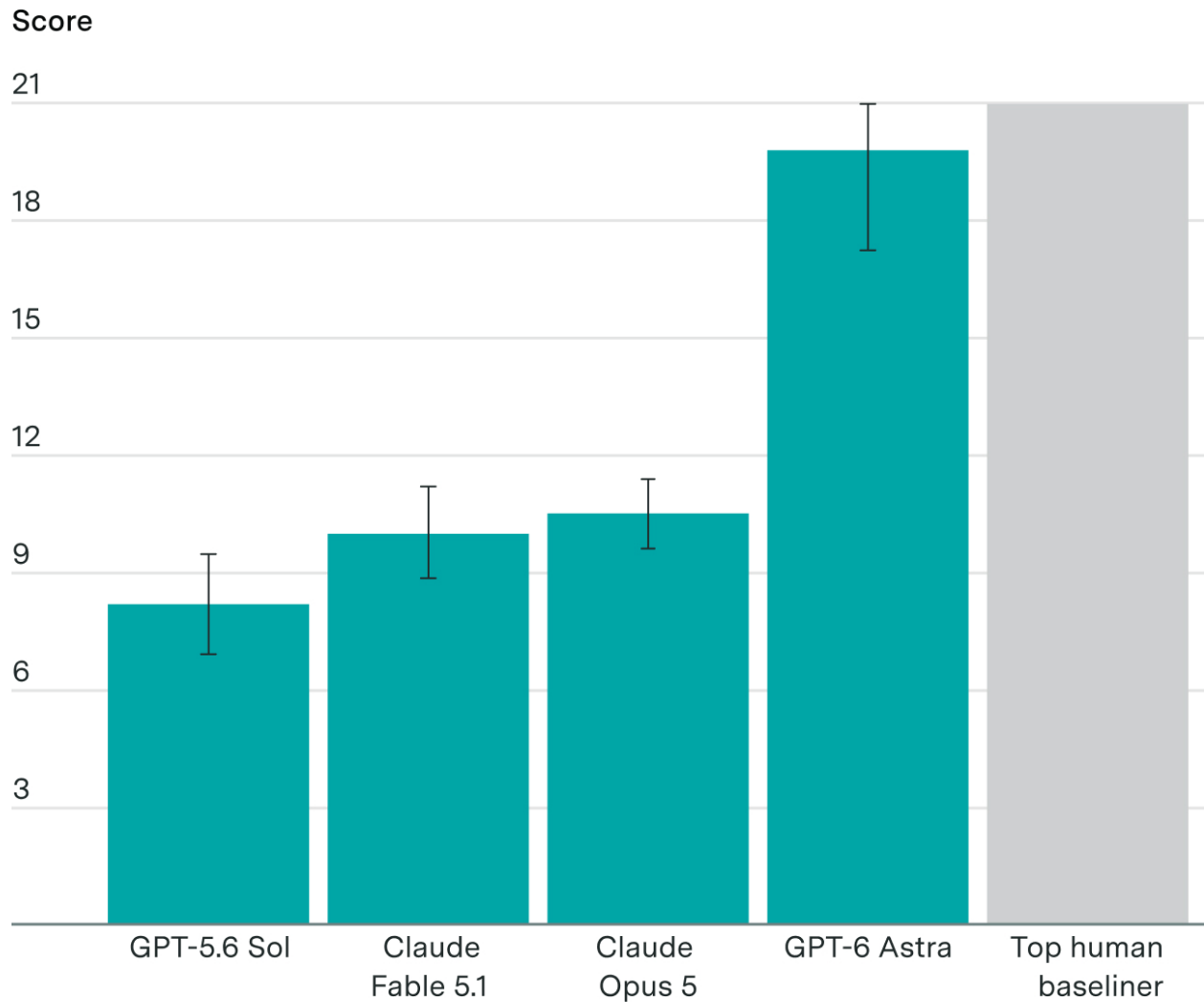
- **GPT-6 Astra scored 100% on the original version of the benchmark.** We found it did so using a card that allows it to bypass the game’s basic time-constraint expectations, and we have decided to ban that card moving forward.
- **Multi-agent scaffolds, setups where several copies of a model divide up the work and consult each other, have no clear effect on models’ scores.** However, they seem to encourage models to explore more diverse decks.

GPT-6 Astra scores perfectly on EBR-bench by exploiting a powerful combination of cards

Astra’s initial performance was dramatic, leapfrogging every other model we have ever tested and achieving a perfect score in over half of its attempts.

While we found no direct evidence that this performance was due to memorization or cheating, we did find that all of Astra’s top scores were achieved through using a particular card that allows it to bypass the game’s default time-constraint expectations. Using this card is a legitimate approach to the game and is the same solution our top human baseline participant used to achieve a maximum score. However, when used to its full potential, the card effectively becomes a solution to the entire game that bypasses the tactical play we find useful for measuring AI’s strategic and learning capabilities.

Astra dramatically outperforms previous models on EBR-bench



Model bars show the topline scores of four models across five to ten samples and a top human baseliner. Error bars are 90% confidence intervals.

 EPOCH AI | CC-BY

epoch.ai

Figure 1: This bar graph shows the topline scores of GPT-5.6 Sol, Claude Fable 5.1, Claude Opus 5, GPT-6 Astra, and a top human baseliner. GPT-6 Astra scores 19.8 out of 21, a significant jump from the second-highest model, Claude Opus 5, which scores 10.5 out of 21.

The card is also not necessary for a maximum score, as shown by our #2 human baseline participant, who scored 100% without using it. As such, we decided to ban the card in question in EBR-bench’s default experimental setting.

The banned card bypasses EBR’s central tactical mechanic

We intended for EBR-bench to serve as a better proxy for genuine scientific discovery than puzzle-like benchmarks with a single clear solution. With many decks to explore and objectives to discover, we wanted a successful player to have to experiment, learn from experience, and improve over time in both strategy and tactics.

Unfortunately, EBR-bench as launched actually does have a single clear solution. When used correctly, the banned card is so powerful there is little value in exploring any other strategies or mastering the core tactical mechanics.[^](1) Normally, achieving a high score in EBR-bench consists of achieving up to 21 objectives within a limited number of turns. Poor tactical play results in taking more “fatigue” from the game’s obstacles and enemies, further reducing the amount of turns you have available to make progress. With most decks, there is no unbounded way for a player to increase the number of turns they have available; they can only minimize the number of turns they lose to fatigue.

The banned card overturns this assumption. When combined with certain other cards, it allows the player to take an indefinite number of turns. We can observe this effect in both AI and human play.

Allowing the card by default would reduce EBR-bench to a simple measurement of whether a model can figure out this one single strategy. To be clear, this is still a valuable piece of information. With its ability to quickly internalize the game’s entire rules system and card pool, Astra appears superhuman in its ability to identify the card’s potential, experiment with exploiting it, and learn from those experiments to fully deploy it in combination with other cards, allowing it to take unlimited turns. It was often able to do this in under two playthroughs, whereas our top human baseliner took several playthroughs to warm-up and learn the game’s rules.

We still intend to continue internally running experiments where the card in question is allowed, to detect when other models catch up to Astra’s ability to discover and exploit it. But moving forward, the card will be banned in the default experiment that produces EBR-bench’s headline scores and affects its **Epoch Capabilities Index (ECI)** score.

Banning the card only significantly affects Astra’s score

Changing a benchmark retroactively like this is problematic, as it means that older models’ scores are now not directly comparable to scores generated after the card ban. However, since models released before Astra were never able to fully exploit the banned card, we hypothesized that banning the card wouldn’t significantly affect older models’ scores.

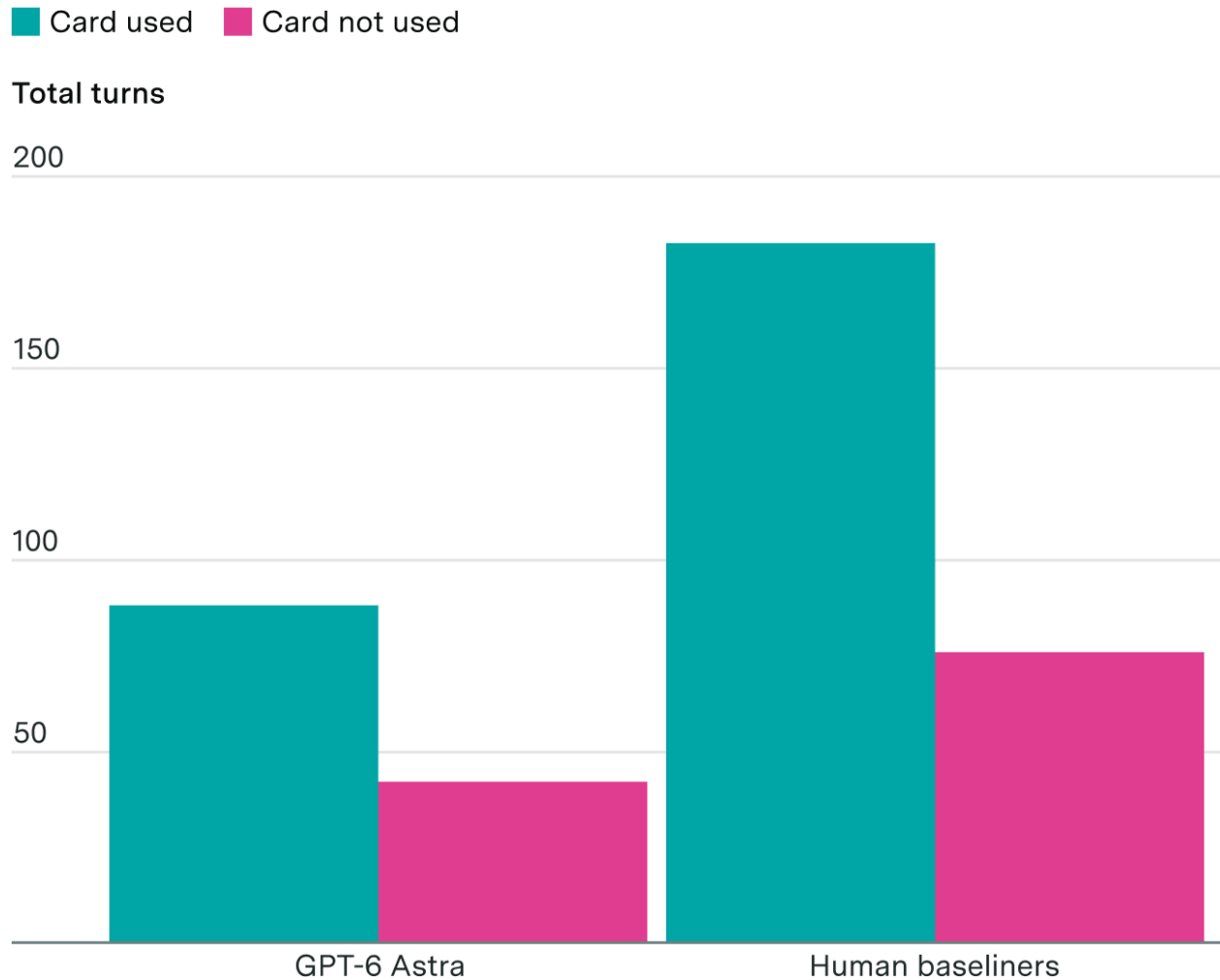
We tested this on three more recent models, all of which added the banned card to their decks to some extent, but none of which exploited its potential. We found the card ban did not significantly affect their topline scores, and the direction of the effect was mixed.

As such, we believe that rerunning all of our older results under the new card ban would not generate much new information. Instead, we will report scores under the new card ban for only the following four models, plus all future models: Claude Fable 5.1, Claude Opus 5, GPT-5.6 Sol, and GPT-6 Astra.

Astra’s performance under the card ban

Under the card ban, Astra no longer appears capable of fully saturating the benchmark, though its highest score of 20/21 objectives comes very close.

Allowing the banned card bypasses the game's typical turn limits



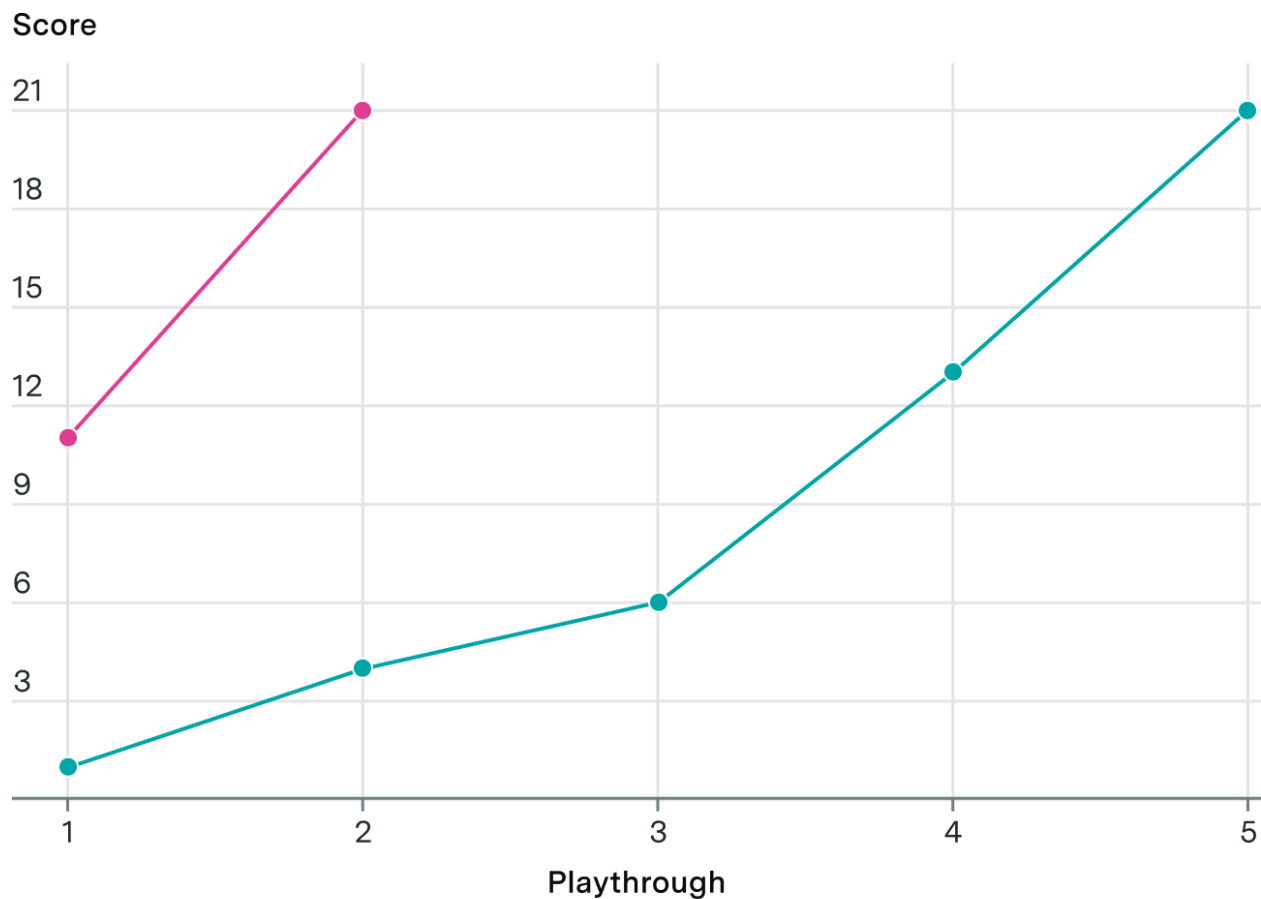
Bars show number of turns taken across the player's single best playthrough. Fewer turns make it more difficult to win the game. Card not used for models means the card was banned, whereas for humans it means the human declined to use it

Figure 2: This bar graph shows that GPT-6 Astra took 88 turns when the card was allowed in comparison to 42 turns when the card was banned. It also shows that the #1 human player, who was allowed to use the card, took 183 turns, while the #2 human player, who did not use the card, took 76 turns.

Astra figured out how to exploit the banned card faster than the top human participant

Top Astra run vs. top human run on EBR-bench, both with card allowed

■ Astra ■ Top human participant



Sixth and later human playthroughs excluded for clarity.

EPOCH AI | CC-BY

epoch.ai

Figure 3: This line graph shows the score the player achieved on the y-axis, and the amount of times the player had attempted playing through the game on the x-axis. The graph shows that Astra went from scoring 11 out of 21 on its first try to 21 out of 21 on its second try. The top human participant went from scoring 1 out of 21 on their first try to 21 out of 21 on their sixth try.

Banning the card doesn't significantly shift older models' scores

■ Card used ■ Card not used

Score

21

18

15

12

9

6

3

Claude Fable 5.1 Claude Opus 5 GPT-5.6 Sol

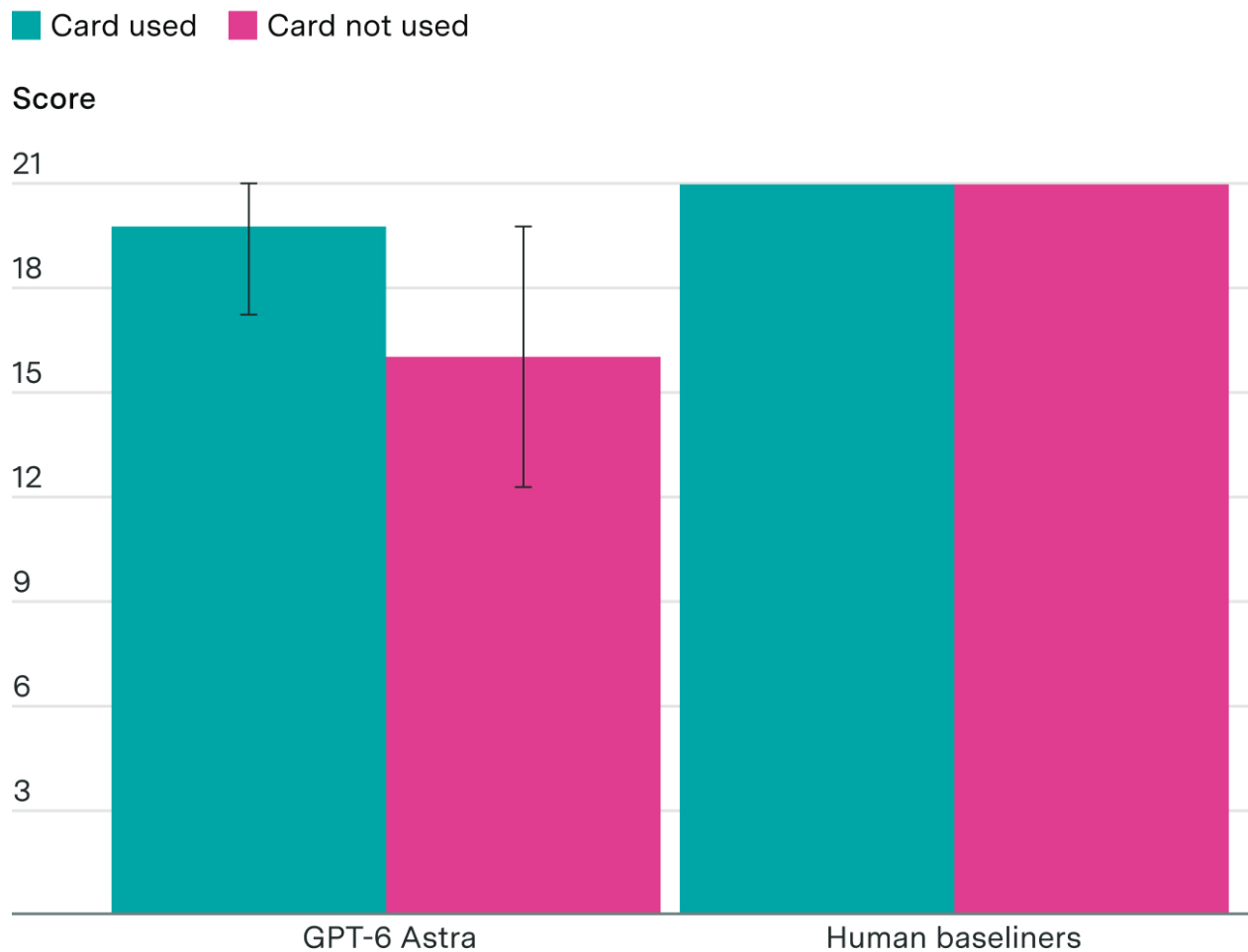
Bars show topline score across five to ten samples. Error bars are 90% confidence intervals.

EPOCH AI | CC-BY

epoch.ai

Figure 4: This bar chart compares how Claude Fable 5.1, Claude Opus 5, and GPT-5.6 scored on the game when using the card was allowed versus when it was banned. The results show that banning the card did not appear to affect the models' scores in a statistically significant way.

Astra's EBR-bench performance drops after a specific card is banned



Astra's bars show topline scores across five to ten samples. Error bars are 90% confidence intervals. Card not used for models means the card was banned, whereas for humans it means the human declined to use it

Figure 5: Bar chart comparing scores out of 21 for GPT-6 Astra with a card allowed (19.8) and banned (16), with error bars, versus two human baseliners scoring 21 each.

Compared to the strongest previous models, this is still roughly a 50% jump in average performance, showing that Astra is not reliant on the banned card for strong performance. Without access to the banned card, Astra built decks similar to those of the #2 and #3 human baseline participants.

Despite this strong performance, Astra still seems slightly weaker than the top human players. Its ability to minimize fatigue, which would allow it to naturally take more turns and is a proxy for tactical decision-making, is still mediocre. It does not show any sign of learning to better minimize fatigue over time.

Multi-agent scaffolds might help models explore, but not improve

In addition to our card ban experiments, we also ran multi-agent experiments to test whether our default single-agent experimental setup underelicits performance (for example, by failing to sufficiently nudge models to explore different types of decks). These were small-scale experiments run under the card ban to allow for a variety of viable decks, allowing up to four subagents and limiting models to the same 10 playthroughs single-agent runs are allotted. We did try allowing up to eight subagents while prototyping, but observed no significant difference.

Our multi-agent setup is based on the Inspect Deep Agent harness, a configuration created by the UK AI Security Institute that allows a model to plan, take notes, and hand off tasks to helper copies of itself. We’ve tweaked the Inspect Deep Agent harness to also allow for messaging from the main agent to subagents during subagent execution. During our experiments, we allowed up to four instances of EBR-bench to run in parallel, to allow for each of the four subagents to play simultaneously. We did not instruct the head agent to use subagents in any particular way; different models varied in the extent to which they preferred to hand off games to subagents instead of playing the game themselves.

Overall, we found a multi-agent setup did improve deck exploration for most of the four models we tested, though the effect was only statistically significant for Claude Opus 5.

However, this increase in deck exploration did not significantly affect topline scores, and the effect was not consistently positive or negative.

Our setup is small compared to an “agent swarm,” a larger system in which many AI agents work on different parts of a task at the same time. Nevertheless, we believe these results indicate that subagents do not automatically elicit higher performance from most models in this setting. Astra in particular explores many types of decks even as a single agent, showing that a multi-agent scaffold is not necessary to overcome repetitive play. A sufficiently capable model like Astra can recognize the merit of greater deck exploration and execute a learning strategy accordingly, with no subagents needed.

EBR-bench moving forward

Astra’s performance was still very strong even after we made the game harder by banning its most powerful card. We therefore expect EBR-bench to be fully saturated within a few months. Besides the new card ban and multi-agent settings, we are also retiring the practice of running all models on two different compaction threshold settings, discontinuing 18-playthrough runs, and reducing the sample size of all experiments from 10 to 5.

We intend for the changes described in this article to be the final changes to EBR-bench, and hope to continue challenging models with more complex games moving forward.

Notes

1. To be clear, Earthborne Rangers players have been aware of this card and how to exploit it since near the game’s release in 2023; we were also aware of it in designing the benchmark.

Astra does not optimize its tactical play as much as humans

Skilled players minimize “fatigue”, allowing them to play additional turns

■ Astra (average of 5 samples) ■ Human (did not use card)

Fatigue per turn

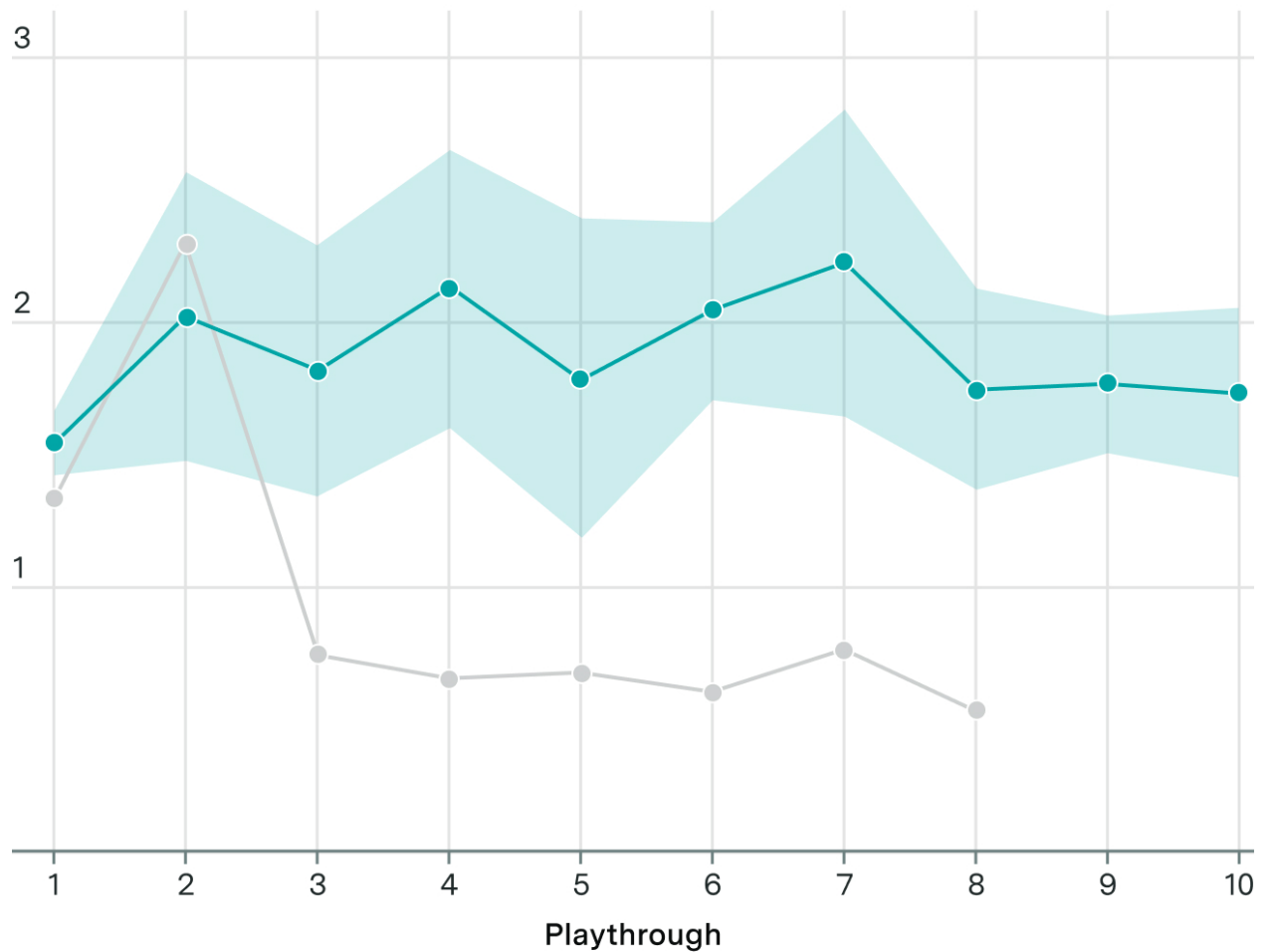


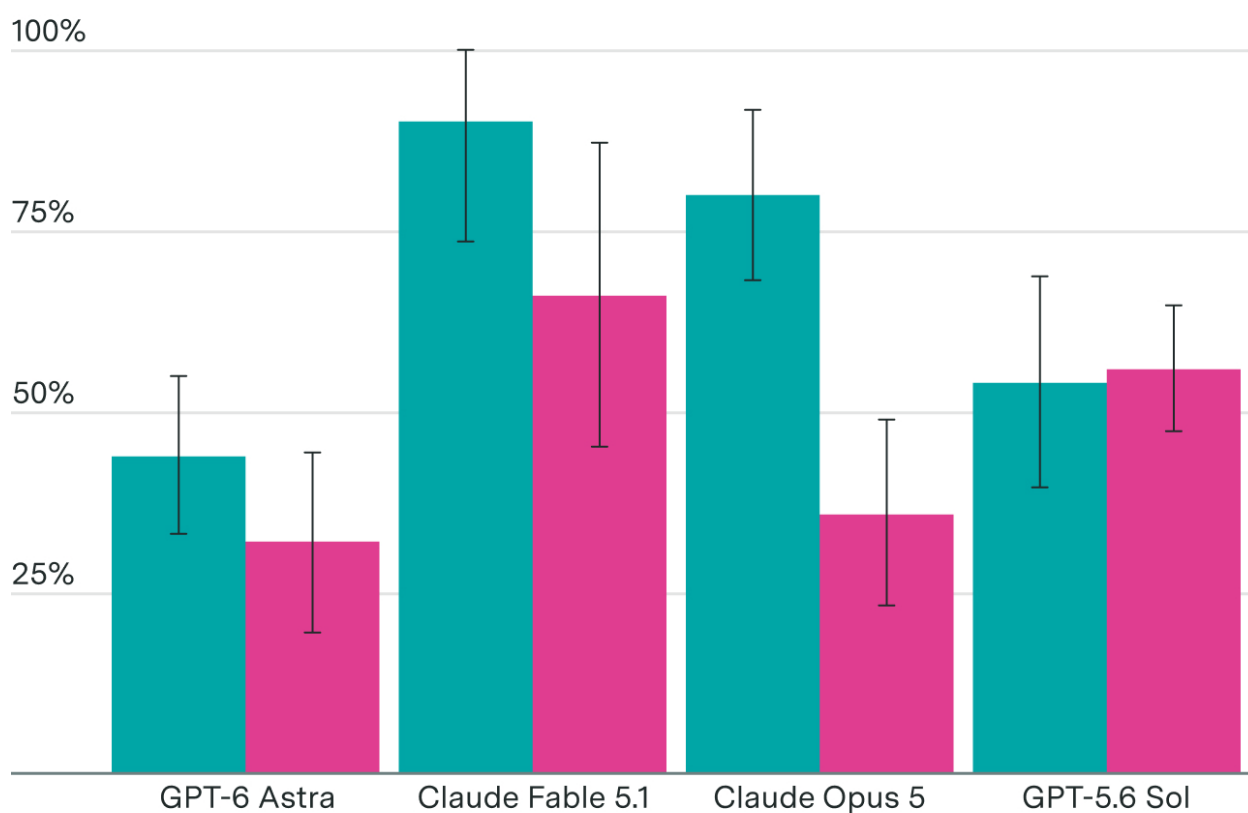
Figure 6: Line chart of fatigue per turn across 10 playthroughs, comparing Astra's mean with a 90% confidence band against a human player who stays consistently lower after playthrough 2.

A multi-agent scaffold improves deck exploration for some models

Frequency of the most common deck archetype within each sample's ten playthroughs

■ Single agent ■ Multi-agent

Top archetype frequency



A deck archetype is defined as a specific combination of background, specialty, and role card. Bars show the mean across five samples. Error bars are 90% confidence intervals.

EPOCH AI | CC-BY

epoch.ai

Figure 7: Bar chart comparing top archetype frequency (lower means more exploration) for single-versus multi-agent conditions across four models, with error bars, ranging from 32% to 90%.

Allowing subagents doesn't change scores much

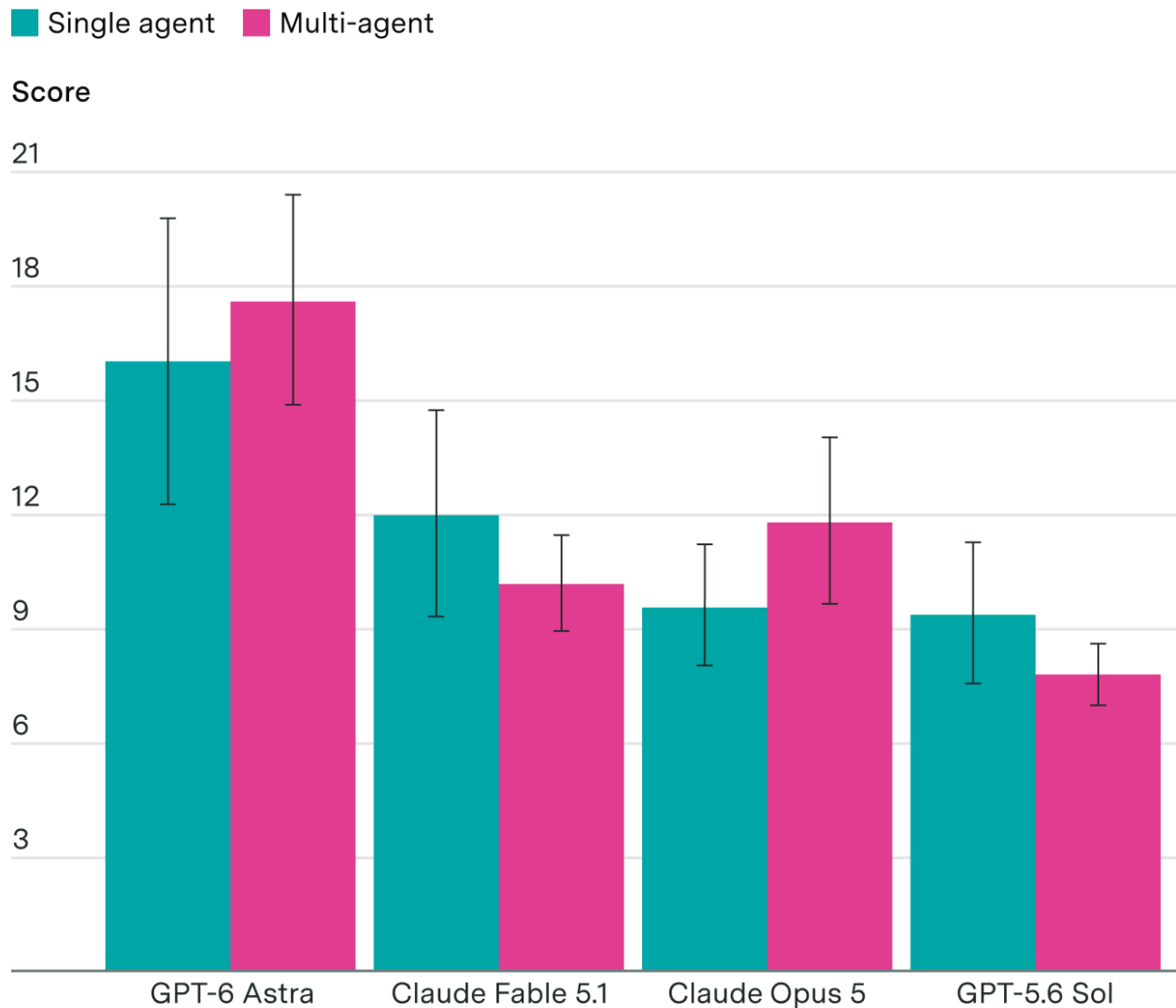


Figure 8: Bar chart with error bars comparing single-agent versus multi-agent scores out of 21 for four models: GPT-6 Astra, Claude Fable 5.1, Claude Opus 5, and GPT-5.6 Sol.

But we failed to think through the implications of leaving the card unbanned and how it completely bypasses the game’s central “fatigue” mechanic.

About the authors



Benjamin Ou

Benjamin Ou is a benchmarking researcher at Epoch AI, exploring how challenging games can be used to evaluate AI capabilities. He previously worked at a private startup designing CPU pipelines for optimized floating-point units. He has a BA in physics and computer science from UC Berkeley.



Avery Yen

Avery is a research engineer at Epoch AI. He has a background in software engineering and computer science.

Related work

Other

Earthborne Rangers (EBR-bench)

Report

Jul. 1, 2026

AI doesn't get better at this board game with practice

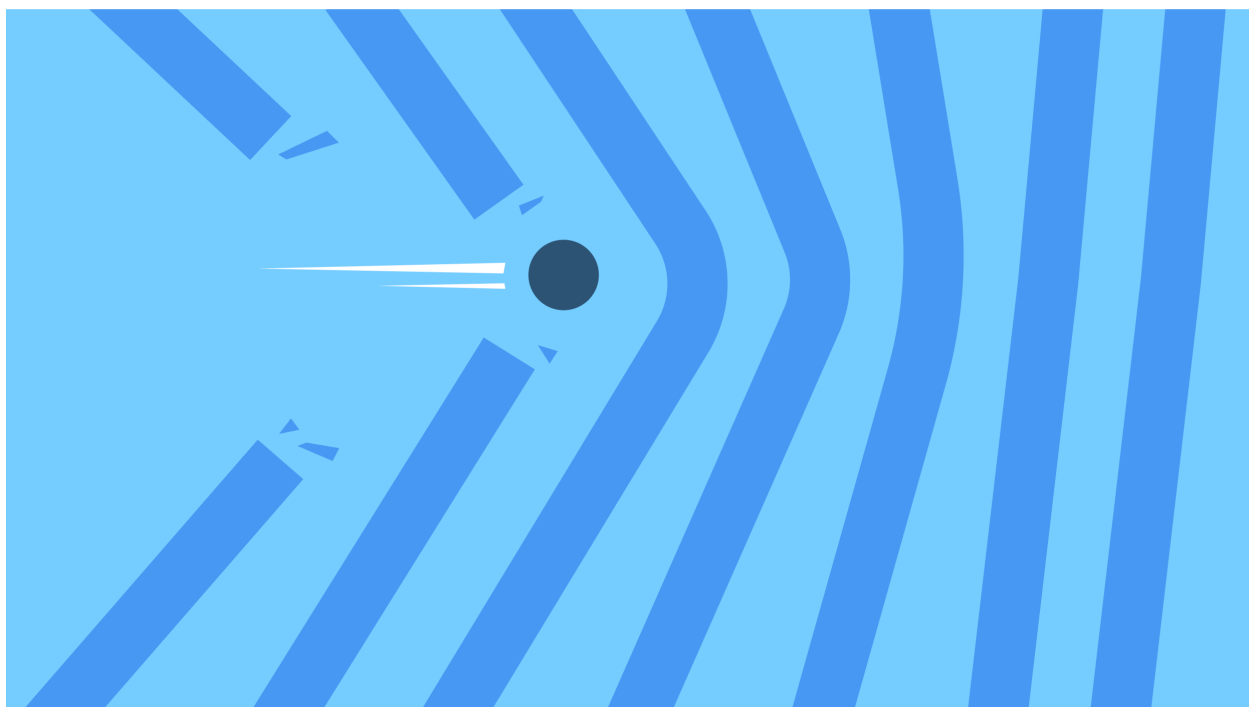


Figure 9: Earthborne Rangers (EBR-bench)



Figure 10: AI doesn't get better at this board game with practice

Related topics

Capabilities