Strategy and structure in Codenames: Comparing human and GPT-4 gameplay

Noah C. Prescott,[1†] Tracey Mills,[2†*] & Jonathan S. Phillips[1]
[1]Dartmouth College, [2]Massachusetts Institute of Technology
[†]Equal contribution

Abstract

Codenames is a rich interactive game that requires players to combine multifaceted conceptual understanding with their capacity to reason strategically about what inferences other players are likely to draw. While prior work has evaluated the extent to which LLMs are able to succeed in playing this game and methods for improving their performance, our approach is rather to investigate and compare the *strategies* and *conceptual structure* humans and a current LLM (GPT-4) employ when playing the game. We find evidence for a shared overarching cognitive strategy of relying on pairwise relatedness judgments to solve Codenames and also show that differences between humans and GPT-4 result from differences in underlying conceptual representations. However, we also find that humans spymasters' strategies reveal a sensitivity to the specific structure of the game that GPT spymasters lack.

1 Introduction

In Codenames, a "spymaster" must solve an open-ended task of generating a one-word clue that will get their teammates–the "guessers"–to select the intended subset of conceptually related words, while avoiding the other unintended words on the board [Chvátil, 2015]. This particular combination of conceptual understanding and strategic coordination with other players makes Codenames a particularly useful target for comparing high-level cognition in human participants and in contemporary LLMs.

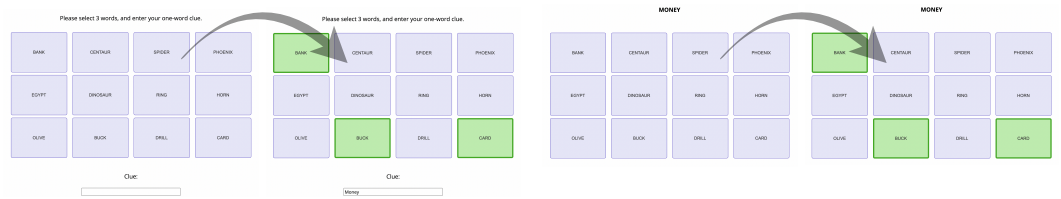
1.1 Related work

Codenames has attracted significant AI interest, with earlier work aimed at evaluating word-embedding based agents [Kim et al., 2019, Jaramillo et al., 2020] which, though successful when paired with models of the same type, are brittle under conceptual misalignment and generate clues that are inscrutable to other model-types or human players [Kim et al., 2019, Jaramillo et al., 2020]. While certain algorithmic augmentations have been suggested to increase their flexibility [Archibald and Brosnahan, 2024, Archibald and Blaylock, 2024] and conceptual alignment [Koyyalagunta et al., 2021], a compelling possibility is that more recent LLMs will naturally circumvent this brittleness by spontaneously incorporating more human-like representations and context-specific adaptive gameplay.

Work aimed at documenting the performance of LLM-based Codenames agents and prompt engineering techniques [Ozturkler et al., 2022, Sidji and Stephenson, 2024] highlights the general success of LLM teams; however, the flexibility of these agents accompanies a lack of interpretability and the cognitive processes underlying patterns in success and failure remain poorly understood. Stephenson et al. [2025] and Hakimov et al. [2025] begin to characterize similarities and differences

*Corresponding author: `temills@mit.edu`

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: CogInterp.



(a) (b)

Figure 1: (a) Example task as seen by spymasters (left) and a completed version of the trial (right). In this illustration, the spymaster comes up with the clue 'money' for words 'bank,' 'buck,' and 'card.' (b) Example task as seen by guessers (left) and a completed version of the trial (right). In this example, the guesser receives the clue 'money' and correctly guesses the target words.

in strategies used by different LLMs including GPT, DeepSeek, Claude, and Llama, and Gemini models. Stephenson et al. [2025] found agents performed best when paired with models of the same type, and exhibited limitations to strategic reasoning in clue-selection, such as selecting clues that over-emphasize a single target word or are similar to non-target words. Hakimov et al. [2025] additionally found that larger models tend to hallucinate less and better follow instructions (e.g. guessing the correct number of words), and that interestingly, word abstractness or frequency did not seem to affect performance.

There is thus still significant uncertainty regarding how LLMs navigate the conceptual and strategic challenges posed by Codenames, and particularly to what extent they approach the game in a way that is similar to or compatible with human players. Here, we pair human and LLM spymasters with human and LLM guessers to understand similarities and differences in play. This approach allows us to identify differences and similarities in the strategies and structure used to play Codenames.

2 Task

To systematically probe the strategies used by humans and LLMs in Codenames, we used a modified version of the original Codenames game [Chvátil, 2015] that isolates key features of interest. Games consisted of one round in which a spymaster selected three of 12 words (randomly selected from the game's original 400-card set) for which to provide a clue, and a guesser selected three of the 12 words to guess based on the clue.

3 Human performance and strategies

Human spymasters were recruited from Prolific ($N = 100$) and completed 10 trials in which they were presented with a board of 12 words and asked to generate a clue and indicate the three target words on the board. Participants were instructed to select targets and generate clues that would give another person the best chance of guessing only the target words. We then randomly selected 400 clue-board pairings from this set, with one clue per unique board, and presented them to a separate group of guessers ($N = 201$). When presented with a board, guessers selected three words that they believed corresponded to the spymaster's intended targets (see Fig. 1). Each guesser saw 20 boards, for an average of 10.02 guessers per board.

Spymasters What cognitive processes do human spymasters use when generating a clue? An obvious starting point would be to generate a clue that is highly conceptually related to each of the three target words while not being highly related to the other words on the board. To test this hypothesis, we recruited additional participants ($N = 77$) to make isolated pairwise relatedness judgments on words and clues. We selected 80 boards from our dataset and created word-clue pairs by pairing each board's clue with each of the 12 words on that board, yielding 960 total pairs. Participants rated the relatedness of 80 word-clue pairs on a 100-point scale, giving us an overall average of 6 relatedness judgments per pair.

Using the average relatedness judgment for each word-clue pair, we found that target words were overall more related to the clue word ($M = 62.82$) than non-target words were ($M = 17.54$, $t(958) = 31.88$, $p < 0.001$). We then tested whether human spymasters' clues also revealed a sensitivity to the specific structure of the game, where each clue should be chosen to be highly related

to exactly three words on the board, but no more. One way to check this is to see if spymasters create a clear conceptual boundary by maximizing the difference in relatedness of the 3rd and 4th most related words on the board. To investigate this, we ranked the words on the board by their relatedness to the clue and normalized each word's relatedness score by dividing it by the sum of all relatedness scores for that clue. We found that the difference (or gap) in relatedness of 3rd and 4th most related words was bigger than the gap between the 4th and 5th most related words (3v4 normalized mean = 0.051; 4v5 normalized mean = 0.025), $p < 0.001$, Cohen's $d = 0.481$. Critically we also found that the 3rd vs. 4th relatedness gap did not differ significantly from the 2nd vs. 3rd relatedness gap (2v3 normalized mean = 0.044), $p = 0.330$, Cohen's $d = 0.109$), suggesting that human spymasters sought to generate clues that were highly related to all three intended target words, but not any of the others. In sum, we find that human spymasters create clues that are conceptually related *specifically* to the three target words.

Guessers We next asked whether human guessers were able to capitalize on the spymasters' strategic clue generation to select the target words given a clue. Overall, we found that individual guessers were successful, guessing 79.21% of words correctly on average, compared to chance performance at 25%. Additionally, guessers got all 3 words correct on almost half of all trials (47.67%), with chance performance at just 0.45%.

4 LLM performance and strategies

To understand what cognitive strategies LLMs employ to play Codenames, and how these differ from humans, we tested GPT-4o-2024-08-06, a popular LLM and one of the most successful Codenames agents in past work [Stephenson et al., 2025], on the same task administered to humans. Using the same 400 boards, we prompted the model with the 12 board words and instructions to select targets and generate a clue. Taking these LLM-sourced clues, we next evaluated the model as a guesser, providing a board and clue and prompting selection of the three words that correspond to the clue. We collected an average of 5.85 independent responses per clue-board pair (complete prompts in Appendix 1).

Spymasters As with humans, we obtained isolated relatedness judgments from the model between LLM-sourced clues and board words for the 80 boards, obtaining 5 independent ratings on a 100-point scale for each of the 960 clue-board word pairings. These ratings were again significantly higher for target words ($M = 70.39$) than for non-targets ($M = 23.63$), $t(958) = 30.66$, $p < 0.001$, suggesting that the model abides by a simple strategy of selecting clues that are most related to target words.

To further examine whether the LLM spymaster selected clues that respected the specific structure of the three-word game, we again ranked the words on the board by their relatedness to the clue and normalized each word's relatedness score. Unlike humans, here we found no clear evidence that the gap in relatedness of 3rd and 4th most related words was bigger than the gap between the 4th and 5th most related words (3v4 normalized mean = 0.036; 4v5 normalized mean = 0.031; $p = 0.366$,

Figure 2: **(a)** Guesser accuracy by condition. **(b)** Difference in mean relatedness of target words to clue vs. difference in guesser accuracy, between human and GPT clues. Human judgments and guessers are in blue, (correlation: $r = 0.55$, $p < .001$, $R^2 = 0.30$) and GPT in orange ($r = 0.38$, $p < .001$, $R^2 = 0.14$).

Cohen's d=0.143). Moreover, we did not find that it differed significantly from the 2nd and 3rd most related words (2v3 normalized mean = 0.044, $p = 0.293$, Cohen's d=0.167). These results collectively suggest that while LLM spymasters sought to select words that were highly related to the target words, the generated clues did not seek to create a conceptual boundary for guessers at exactly three words.

Guessers Partnerships between LLM spymasters and LLM guessers were generally successful, with guessers guessing 79.24% of words correctly on average and guessers getting all 3 words correct on 43.49% of trials (with chance performance at 25% and 0.45% respectively).

5 Comparing humans and LLMs

To directly compare human and LLM performance across all task components, we conducted a full cross-evaluation design. Maintaining the same experimental setup, we presented the model with human-generated clues from the initial experiment and collected new model guesses. The model also provided relatedness judgments for all human clue-word pairs. Conversely, human participants recruited from Prolific ($N = 117$) made guesses on the GPT-generated clues using identical procedures to the initial human guessing task. Furthermore, we recruited additional human participants from Prolific ($N = 119$) to judge the relatedness of LLM clue-board pairs following the same protocol used to obtain relatedness judgments for human-generated clues.

Performance To examine clue generation and interpretation in human-AI collaboration, we analyzed performance across four pairwise combinations of human and LLM spymasters with human and LLM guessers. We specifically targeted guesser accuracy, measured as the proportion of correctly identified target words, across conditions. When guessing targets for human-generated clues, humans ($M = 0.79$, $\sigma = 0.149$) and GPT ($M = 0.80$, $\sigma = 0.192$) performed similarly well ($t = -0.646$, $p = 0.518$, Cohen's $d = -0.046$). However, human participants had significantly reduced accuracy when guessing targets for GPT-generated clues ($M = 0.74$, $S\sigma = 0.165$) compared to human-generated clues ($t = -4.552$, $p < 0.001$, Cohen's $d = -0.322$). Conversely, the LLM's guessing performance was not significantly affected by clue source, performing similarly on both LLM-generated ($M = 0.79$, $\sigma = 0.175$) and human-generated clues ($M = 0.80$, $\sigma = 0.192$; $t = -0.4053$, $p = 0.685$, Cohen's $d = -0.0287$); see Fig. 2a. When presented with LLM-generated clues, human participants performed significantly worse ($M = 0.739$, $\sigma = 0.165$) than the LLM did ($M = 0.793$, $\sigma = 0.175$; $t = -4.4186$, $p < 0.001$, Cohen's $d = -0.3124$).

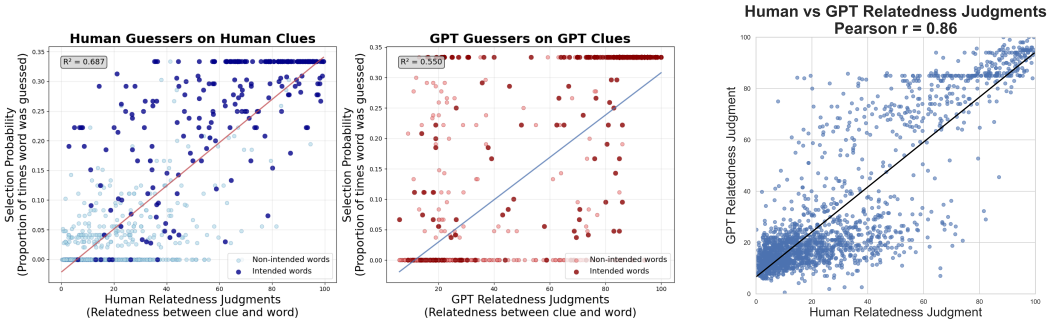


Figure 3: **(a)** Word relatedness to clue vs. probability of being guessed, for humans (left) and GPT (right). Each point represents a single clue-word pair, with the x-axis indicating the average relatedness judgment for that pair, and the y-axis indicating the proportion of participants who guessed the word when cued with the clue. Human relatedness judgments explain 68.7% of variance of selection of human guesses on human clues (total words = 960, $r = 0.8288$, $p < 0.001$). LLM relatedness judgments explain 55.0% of the variance of selection of LLM guesses on LLM clues (total words = 960, $r = 0.7419$, $p < 0.001$.) **(b)** Human vs. GPT relatedness judgments. Each point represents a single clue-word pair, with the x-axis indicating the average human relatedness judgment for that pair, and the y-axis indicating the average GPT relatedness judgment for that pair.

Cognitive strategies What drives differences in human guesser performance on human and GPT-generated clues? The probabilities of selecting a guess for both human participants and the LLM are predicted by their respective relatedness judgments (humans: $r = 0.83$, $p < 0.001$; LLM: $r = 0.74$, $p < 0.001$; Fig. 3a), illustrating a shared overarching strategy employed for the task. One possible explanation for the difference is conceptual misalignment between humans and the LLM, however, the correlation between human and GPT relatedness judgments is relatively high ($r = 0.86$, $p < 0.001$; see Fig. 3b).

Another possibility is that when humans perform better on human than GPT-generated clues, GPT spymasters and guessers employ a different strategy than humans. On these boards, however, GPT guessers also prefer human clues (83.4% vs 75.6% accuracy, $t = 4.675$, $p < 0.001$), and human and GPT guesser accuracy is predicted by relatedness judgments to a similar extent on these (n = 43, human: $r = 0.48$, $p < 0.001$, GPT: $r = 0.33$, $p = 0.032$) as on all boards (n = 80, human: r = 0.5469, p < 0.001, GPT: r = 0.38, p < 0.001), suggesting a consistent strategy across boards. Indeed, differences in performance on human and GPT clues are predicted simply by differences in clue-target word relatedness (Fig. 2b). Human guessers' preference for human-generated clues is therefore probably best explained by greater conceptual relatedness between human clues and target words.

6 Limitations and open questions

This work highlights both similarities and differences in the cognitive strategies used by humans and LLMs when playing a modified version of Codenames [Chvátil, 2015]. At the same time, our results and interpretation are limited in a number of key ways. First, we compared human players to a single closed-source LLM model (GPT-4o-2024-08-06), and so we are limited in our ability to investigate why some of these patterns arose and cannot be sure that they will generalize to other LLMs. Second, we did not conduct a wide range of prompting techniques, and so it remains possible that the key differences between the human and LLM gameplay could be ameliorated with different prompting techniques. In addition, the present work raises but does not resolve a number of intriguing questions, including why human guessers prefer human-generated clues while LLMs do not differentiate between human- and LLM-generated clues. Relatedly, an important goal is to connect the observed patterns to a deeper understanding of the difference in conceptual representation between humans and LLMs. For example, recent work has pointed out that conceptual representation in LLMs is more optimally compressed than in humans, but also results in a loss of nuance Shani et al. [2025], and it would be useful to know whether some of our findings can be connected to this phenomenon (e.g., LLM relatedness ratings are more bimodal than humans as can be observed in Fig. 3b). At the same time, we suspect that some of the key findings, especially those related to differences in strategy (e.g., the relatedness gap between the 3rd and 4th most related words), are unlikely to be explained by broad differences in conceptual representation.

7 Conclusion

In sum, we have demonstrated that human and GPT Codenames spymasters and guessers use similar cognitive strategies based on pairwise relatedness judgments, and are overall conceptually aligned. However, human guessers systematically preform better when given human-generated rather than GPT-generated clues. We identified two independent explanations for this. First, differences in human performance on human and GPT clues are predicted by differences in relatedness judgments, suggesting that humans are better able to generate clues that are highly conceptually related to target words. Second, human spymasters exhibit a strategic sensitivity to specific game rules by maximizing the relatedness gap between the 3rd and 4th most related words, whereas LLMs spymasters exhibit no such strategic boundary formation.

We hope future work builds on these findings and employs a wider range of open-source models to better understand the similarities and differences between human and LLM gameplay in cognitively rich interactive games like Codenames.

References

Christopher Archibald and Diego Blaylock. Noisy communication modeling for improved cooperation in codenames. In *2024 IEEE Conference on Games (CoG)*, pages 1–8. IEEE, 2024.

Christopher Archibald and Spencer Brosnahan. Adapting to teammates in a cooperative language game. *arXiv preprint arXiv:2403.00823*, 2024.

Vlaada Chvátil. *Codenames*. Board Game, 2015. Designer: Vlaada Chvátil. Publisher: Czech Games Edition. Year Released: 2015.

Sherzod Hakimov, Lara Pfennigschmidt, and David Schlangen. Ad-hoc concept forming in the game codenames as a means for evaluating large language models. *arXiv preprint arXiv:2502.11707*, 2025.

Catalina Jaramillo, Megan Charity, Rodrigo Canaan, and Julian Togelius. Word autobots: Using transformers for word association in the game codenames. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 16, pages 231–237, 2020.

Andrew Kim, Maxim Ruzmaykin, Aaron Truong, and Adam Summerville. Cooperation and codenames: Understanding natural language processing via codenames. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 15, pages 160–166, 2019.

Divya Koyyalagunta, Anna Sun, Rachel Lea Draelos, and Cynthia Rudin. Playing codenames with language graphs and word embeddings. *Journal of Artificial Intelligence Research*, 71:319–346, 2021.

Batu Ozturkler, Nikolay Malkin, Zhen Wang, and Nebojsa Jojic. Thinksum: Probabilistic reasoning over sets using large language models. *arXiv preprint arXiv:2210.01293*, 2022.

Chen Shani, Liron Soffer, Dan Jurafsky, Yann LeCun, and Ravid Shwartz-Ziv. From tokens to thoughts: How llms and humans trade compression for meaning. *arXiv preprint arXiv:2505.17117*, 2025.

Matthew Sidji and Matthew Stephenson. Prompt engineering chatgpt for codenames. In *2024 IEEE Conference on Games (CoG)*, pages 1–4. IEEE, 2024.

Matthew Stephenson, Matthew Sidji, and Benoît Ronval. Codenames as a benchmark for large language models. *IEEE Transactions on Games*, pages 1–12, 2025. doi: 10.1109/TG.2025.3563810.

A Technical Appendices and Supplementary Material

All data and code are available at: https://github.com/NoahPrescott/Codenames-Strategy-and-Structure

A.1 Supplementary methods

A.1.1 Model prompts

When obtaining clues and target words from GPT spymasters, we prompted the model as follows for each board: "You will be shown a board of 12 words. Your task is to select 3 of the words, and come up with a one-word clue that would allow another player to guess those words. After you choose the 3 words and give your clue, another player will be shown the same board of 12 words and your clue, and will be asked to guess the 3 words that you had in mind. Your clue should apply to each of the 3 words you chose more than it applies to any of the other words. The words are: ", followed by the list of words.

When obtaining guessed words from GPT guessers, we prompted the model as follows for each board: "I am going to give you a one-word clue, along with a list of 12 words. I chose the clue to help you guess exactly 3 of the words in the list. Your task is to list the 3 words that you think I have in mind based on the clue. Please simply list the 3 words that you think I have in mind based on the clue" followed by the clue and board words.

When obtaining isolated relatedness judgments between clues and board words from GPT, we prompted the model as follows: "I am going to provide you with 2 words. Your task is to assess them on a relatedness scale of 1 to 100, 100 being very related and 0 being not related.", followed by the two words, and then "Please just provide your score."