SEARCH AND HUMAN PERCEPTION OF CREATIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

We present empirical evidence that human perception of creativity is correlated with how difficult the search problem for the creative solution is. We use a large dataset of high-quality annotations of chess moves as "brilliant", as well as present a novel dataset of human reactions to Wordle games. In both the analysis of expert-annotated brilliant moves in chess and the preliminary analysis of Wordle, we show that when a play requires a particular shape of search tree, it is more likely to be perceived as brilliant or creative, over and above the quality of the move in the case of chess. Although human creativity is complex and multifaceted, our work indicates that the mathematical structure of the search problem being solved is likely part of what we think of as creativity, across at least two domains.

1 INTRODUCTION

Artifacts "created" by generative AI are sometimes perceived as creative despite the fact that they can be merely the result of an algorithmic process. However, as argued by Hertzmann (2018) and Chiang (2024), it is plausible that a person thinking that an AlphaGo move is creative is simply making a mistake: if art and creativity are essentially about artists making choices in order to communicate something to humans, artefacts that are merely the result of computation are not the result of creativity and are not art. Without wholly disagreeing with this perspective, this paper aims to empirically show that human perception of creativity is in part mathematical: it is related to the mathematical properties of the search process for the artefact.

In this paper, we present empirical evidence that human perception of creativity is related to the nature of the search problem for the creative solution, providing evidence that what humans conceive of as creativity is, at least in part, about the mathematical structure of the problem, building on a long history of theory along those lines (Boden, 1998; Colton et al., 2011).

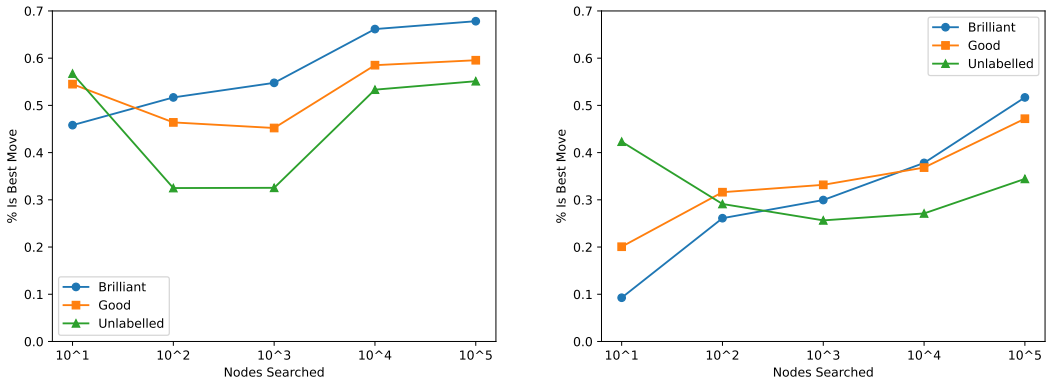
We expand on prior work on human perception of creativity in chess (Zaidi & Guerzhoy, 2024) by using a much higher-quality dataset of chess move annotations by experts as well as novel analyses, and confirm that human perception of creativity in chess is correlated with game tree "shape" and depth and posit those are related to creativity. We also present a novel dataset of human perception of creativity in Wordle games, and show that an analogous finding holds.

2 BACKGROUND

2.1 CHESS

Chess aesthetics research goes back decades (Margulies, 1977; Walls, 1997; Levitt & Friedgood, 1995; Iqbal, 2006) and mostly focuses on chess puzzles and deriving general non-black-box heuristics for chess aesthetics. Zaidi & Guerzhoy (2024) present a machine learning model that predicts human perception of creativity in chess moves by predicting "brilliant" move annotations in a dataset of Lichess Studies (Lichess, 2025), a dataset of chess games annotated by expert and non-expert game users. By using features derived from the game trees associated with each move, Zaidi & Guerzhoy (2024) show that it is possible to predict human perception of creativity in non-experts with an accuracy of 79% with a 50% chance baseline. They also show that it is possible to predict human perception of creativity by comparing the evaluation of a stronger (looks further ahead) and a weaker chess engine of the position, suggesting that perception of creativity is related to the size of the game tree that was likely explored.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Using LC0 weights

(b) Using Maia weights

Figure 1: Percent of moves in database where the engine believes the move is the best move, conditioned on move type. Brilliant moves (blue) show the largest evaluation gain as search depth increases, indicating they are 'hidden' from shallow search.

In a somewhat related line of work, Jenner et al. (2024) have shown that Transformers that play chess use lookahead by pointing out that attention patterns correspond to potential future moves.

2.2 WORDLE

Wordle (Wardle, 2021) is a popular online game that can also involve search: players can strategically play a guess to explore and maximize their chances of winning more quickly.

Luo et al. (2025) have shown that it is possible to predict human perception of amusement at Wordle games from game features alone. Liang et al. (2024) have shown that human gameplay is biased towards words with certain semantic, orthographic, and phonological properties, and those deviations were shown by Luo et al. (2025) to be connected to human perception of amusement – exploring the search space for guesses more aggressively is correlated with amusement reactions. Luo et al. (2025) use reactions on the subreddit `r/wordle` to obtain human amusement reactions to Wordle games. In this work, we use a similar technique to find Wordle games that humans perceive as creative.

3 EXPERIMENTS: CHESS

3.1 DATA

This work follows up on Zaidi & Guerzhoy (2024) by using a new, high-quality dataset of chess moves annotated by experts. The dataset is based on the Chess Informant Quality Base 2024 (Šahovski Informator, 1966–), a dataset of chess games played by strong players and annotated by experts. The dataset contains moves annotated as "brilliant" (marked "!!"), "good" (marked "!"), and unlabeled (neither brilliant nor good). Zaidi & Guerzhoy (2024) previously used a dataset annotated by a mix of experts and non-experts which was scraped from Lichess.org through their "Study" feature. The data from Lichess had several shortcomings: Lichess studies sometimes analyze openings and puzzles or synthetic examples rather than real games, anyone can annotate any game regardless of skill which introduces greater variance in what is deemed brilliant, and there aren't as many brilliant moves labelled in the original dataset.

After processing, the final Chess Informant dataset contained 5278 brilliant moves, 8591 good moves, and 6192 unlabeled moves. Below, all training and metrics use class balancing with 50% weight towards brilliant moves, and 25% weight towards each of good and unlabeled, summing to 50% weight on non-brilliant moves.

3.2 Feature Engineering

Tree features are generated for each move as in Zaidi & Guerzhoy (2024). For each move, search trees are expanded starting from the board state before the move (the parent board) using the Leela Chess Zero engine (Anderson & Meyer, 2022). Expansion is repeated for two sets of neural network weights: T82-768x15x24h-swa-7464000, obtained through Lc0 self-play training, and Maia-1900, which was trained to play similarly to a 1900 ELO human. For each weight, 5 trees are generated with varying numbers of nodes searched: $10^1, 10^2, 10^3, 10^4, 10^5$. This totals to 10 search trees generated per move.

For each search tree, 2 boolean features are calculated: is the move in the tree, and is the move the best move. Then, 22 features are used to encode the characteristics of the search tree, capturing information about the shape and how the engine evaluates the move strength. These 22 features are also computed for each subtree rooted at the children of the parent board.

Since the number of children are variable (depending on the number of possible moves evaluated), the subtrees are grouped and their features aggregated to yield a constant number of features. To preserve information about the move being evaluating, the subtree following that move is set aside as its own group, so its features are not aggregated. There are 4 other groups: children moves that increase or decrease the predicted win chance relative to the parent board, and children moves that have a positive or negative predicted win chance (0 is draw). The features across these 4 subtree groupings are aggregated using 4 aggregation functions: mean, standard deviation, min and max.

This yields: $(2 + 22)$ features/tree $+ 22$ features/subtree $\times$ $(1$ subtree after move being evaluated $+$ 4 subtree groups $* 4$ aggregation functions$) = 398$ features/tree. Since there are 10 search trees per move, this yields 3980 total features per move.

4 Results: Chess

The neural network architecture described in Zaidi & Guerzhoy (2024) (AggReduce) was trained, using 5-fold CV grid search to identify optimal hyperparameters. For this data, the optimal hyperparameters were h1=50, h2=600, h3=50, and num epochs=6.

4.1 High-Quality Chess Dataset Performance

Both the new neural network trained on the Chess Informant data and the prior model trained on Lichess data were used to predict classes on the test data from both datasets. These results are reported in Table 1.

The neural network trained and evaluated on the new Chess Informant data achieves 85.1% accuracy, with a true positive rate of 87.2%, true negative rate of 83.1%, positive predictive value of 83.7%, and negative predictive value of 86.6%. Nearly all of these metrics are higher than the Lichess-trained neural net evaluated on the Lichess dataset (save for true negative rate, which was slightly higher). The accuracy on good moves is 72.0%, and on unlabeled moves is 94.1%. This validates that, as one may expect, good moves are more likely to be incorrectly classified as brilliant than unlabeled moves.

The model trained on Chess Informant data gets 75.4% accuracy on Lichess data, while the model trained on Lichess gets 68.7% accuracy on Chess Informant data, with a low true positive rate of 53.1%. Thus when trained on the new data, the model generalizes better on out-of-distribution data.

Table 1: Chess Informant (CI) and Lichess (LiC)-trained model predictions on CI and LiC data, shown as class-balanced percentages of the total dataset.

Dataset	**True Label**	**CI Model Pred.**		**LiC Model Pred.**	
		Brill.	**Not Brill.**	**Brill.**	**Not Brill.**
CI Data	Brilliant	43.6%	6.4%	26.5%	23.5%
	Good	7.0%	18.0%	4.7%	20.3%
	Unlabeled	1.5%	23.5%	3.1%	21.9%
LiC Data	Brilliant	42.1%	7.9%	35.5%	14.5%
	Good	13.2%	11.8%	5.3%	19.7%
	Unlabeled	3.5%	21.5%	1.8%	23.2%

4.1.1 BRILLIANCE VS BEST MOVE

As search depth is increased, and when using the LC0 weights over Maia weights, a brilliant move is more likely to be ranked as the best move (highest win chance according to LC0), as shown by comparing Figures 1a and 1b. But even with the strongest analysis (10^5 moves searched, LC0 weights), only 67.8% of brilliant moves are ranked as the best move. Furthermore, good and unlabeled moves are more often than not also the best moves available; with the strongest analysis, 59.6% of good moves and 55.1% of unlabeled moves are the best move. This validates that brilliance is not solely dependent on if a move is the best one available at a given time, in fact it is not particularly informative alone.

4.1.2 MOST INFORMATIVE INDIVIDUAL FEATURES

Logistic regression models trained on individual features were ranked from highest accuracy to lowest accuracy. The top 5 features, shown in Table 2, looked at the branching factor, maximum tree width, width at a depth of 2, and the mean width across all the tree depths. These all describe the shape of the search tree. In fact, the first feature not explicitly related to tree shape is at rank 15: is the move found in the tree or not (using Maia weights with 10^1 nodes searched). The model trained on this feature yielded test accuracies of 80.3% on brilliant moves, 34.7% on good moves, 69.5% on unlabeled moves, and 66.2% overall, indicating it struggles to correctly identify good moves as non-brilliant. Based on these results, when looking at individual features, the shape of the search tree can be informative for distinguishing brilliant moves from non-brilliant ones, and might be more informative than features like move strength.

The subtree root and direction (Dir) in Table 2 also show some characteristics regarding the shape of brilliant move search trees. Note, Dir means the direction the feature must move to be classified as brilliant. The model finds a move more brilliant when the branching factor is lower for the post-move subtree (the subtree rooted at the board after playing the move of interest). This is also true for the max width, width at depth 2, and mean width. This means that brilliant moves are more likely to lead to thinner trees following the move. At the same time, the model finds a move more brilliant when the max width of the parent tree is higher. Combining these insights, a brilliant move may be one which has a wide search tree before the move, but a thin one following the move, suggesting brilliance in chess may partially be finding the narrow path to victory among many possible paths.

Table 2: Features achieving the highest class-balanced accuracy (total) when used to train logistic regression models. A feature is defined by the subtree feature and which subtree it was computed for (which is defined by number of nodes, subtree root, and weights used). All features here use Maia.

			Accuracy				
Rk	**Nodes**	**Feature / Root**	**Brill.**	**Good**	**Unlab.**	**Tot.**	**Dir**
1	10^2	Branch. Factor / Post	.632	.659	.837	.690	Lo
2	10^2	Max Width / Post	.709	.536	.770	.681	Lo
3	10^2	Width@Depth2 / Post	.797	.424	.703	.680	Lo
4	10^5	Max Width / Parent	.651	.619	.750	.668	Hi
5	10^2	Mean Width / Post	.820	.376	.651	.667	Lo
15	10^1	Move In Tree / Parent	.803	.347	.695	.662	Lo

4.1.3 BRILLIANT MOVES ELUDE WEAKER ENGINE

To examine how the predicted strength of a move changes based on engine strength, Figure 2 shows predicted win chance as the number of nodes searched is increased, for each weight and move type. We find that brilliant moves are often judged as having lower win chance by Maia, and a higher win chance by the LC0 weights. Brilliant moves also exhibit the largest gain in predicted win chance as the engine searches more nodes. Based on these observations, weaker engines (those using Maia weights and/or smaller search depth) may struggle more to discover why a brilliant move is strong, and instead think they are weak.

To explore this further, the dataset was modified such that for the trees explored with Maia, the post-move predicted win chance was set to -1. This would mean Maia thinks the move was very

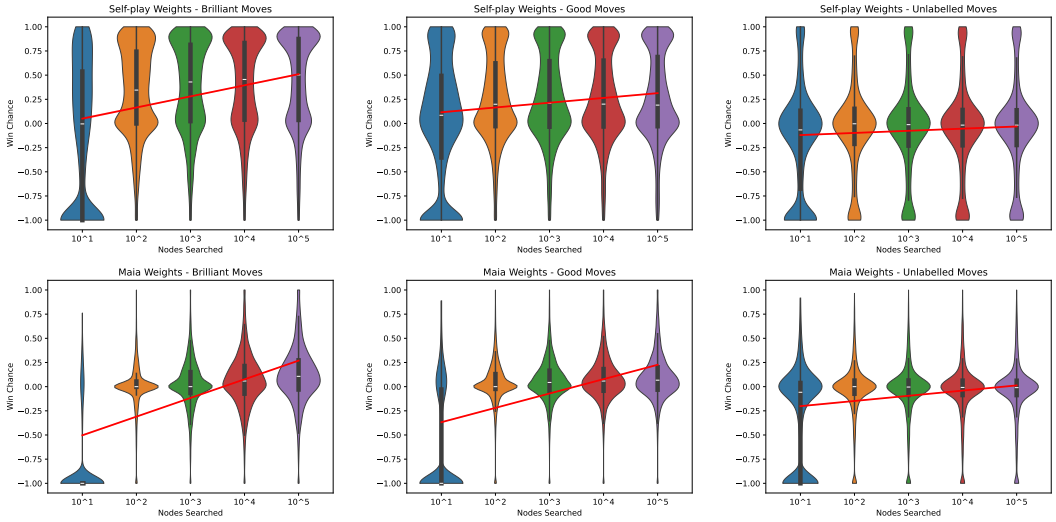


Figure 2: Predicted win chance vs nodes searched using LC0 and Maia weights, by move type. Brilliant moves (left) show the largest gain in predicted win chance as search depth increases, indicating they are 'hidden' from shallow search.

bad, believing the player will certainly lose the game after making the move. After making this adjustment, the brilliance score went up by around 30.2% on average. For already brilliant moves, the average score went up 33.3%, for good moves it went up 41.0%, and for unlabeled moves it went up only 13.2%. This indicates that when the weaker engine believes the move is bad, the move is believed to be more brilliant all else being kept constant. This effect is most pronounced for good moves, indicating it might be a key difference between brilliant and good moves, and the effect is least pronounced for unlabeled moves, which is sensible since Maia thinking the move is weak doesn't mean anything about brilliance if other traits of brilliance aren't shown. For example, an unlabeled move might truly be weak or neutral.

5 Experiments: Wordle

In this section, we present a novel dataset of human perception of creativity in Wordle games, and show that human perception of creativity is correlated with the search problem being solved. We scrape a large number of Wordle games and reactions to them from the `r/Wordle` subreddit, use GPT-3.5 Turbo to find human reactions that express perception of creativity and then hand-verify them, and correlate human perception of creativity with the shape of the search problem being solved.

5.1 Dataset

`r/Wordle` is a subreddit dedicated to the Wordle game, with Reddit users posting their games and commenting on their own and others' games. We scrape the subreddit using the Pushshift API to obtain posts and comments related to Wordle games. We then use GPT-3.5 Turbo to find comments that express perception of creativity, and hand-verify them.

We collected 25,000 games, then used GPT-3.5 Turbo with two prompts to find candidates for comments expressing perception of creativity. The system prompts were:

> **Prompt 1:** "You are a helpful assistant evaluating whether a comment expresses that the original Wordle gameplay (in the post it replies to) was creative. You are not judging whether the comment itself is creative. Instead, you are judging whether the comment perceives the original gameplay to be creative. Look for words or phrases that express admiration for strategy, deduction, cleverness, or creativity (e.g., 'smart guess', 'great strategy', 'clever', 'genius', 'you really thought this through'). Give low scores to comments that do not respond meaningfully to the original gameplay or do not mention creativity, strategy, or deduction."

Comment	Creativity Score (/10)	Game	Answer
In a strategic way. CLK are all very common consonants that makes sense to guess in your first two guess esp with NS already being eliminated. Same with AU being the only vowels left after OIE being gone....	9	N O I S E C A U L K	caulk
Best result so far! My choice of the first word was really helpful. Then considered SPIRE as #2, before opting for the correct answer instead.	9	S O A R E S H I R E	shire
I'd suggest swapping a few consonants for more common ones. Based on your strategy to hit all vowels in three words, maybe FLOAT-CRUDE-SHINY? Edit: Or TONAL-CRIED-MUSHY.	9	D O G M A E R U P T S H I N Y S T O R Y	story
that is really really really fricking smart	9	D E A R Y O I N K S C O U G H J U M B O F L O U T	flout
I usually play with the hard mode rules, but I've stopped bothering in situations like this. It takes more strategy to find a word that eliminates multiple possibilities than it does to just brute-force guesses.	9	S T A R E P I N C H S C A R E	scare
Been forever since I got a three turn. I always start with those first two words as it gives me all five vowels as well as some big letters (MNTS). Seeing I had the N in the right spot and only an I had me stumped at first but I worked through the word and letter patterns I knew...	9	A B O U T M I N E S C Y N I C	cynic
The 6 words remaining after COULD all start with the same 3 letters. Guessing CHEEK, if it's wrong, can eliminate 2 other possible words, CHEEP and CHECK. Guessing CHEEP or CHECK can only eliminate CHEEK, and guessing any of the other words can only eliminate that word.	9	A R I S E C O U L D C H E W Y C H E C K C H E E K	cheek
After my first two openers, I realized the word was gonna end in "OUND," probably ROUND, FOUND, or BOUND. So I figured I should submit a guess that includes R F and B. I think I was pretty successful at that.	9	P I A N O M U L E S B A R F S F O U N D	found
OK. That's how I try to play it unless I'm just plain stumped. And I'll give it a good 20 minutes, walk away for a while, come back and try again before making an obvious wrong guess.	8	S T O L E A M P L E M A P L E	maple
I narrowed it down to RAKER, FAKER, BAKER. Not sure if the first one is really a valid answer but didn't want to chance it...so if my fifth guess came up all grey that's what I was going for.	8	D E A L T C R E A M P A Y E R W A K E R B U F F Y B A K E R	baker

Figure 3: Sample comments from `r/Wordle` expressing perception of creativity in Wordle games. Games with scores of 8–9 are shown. Games are rendered with standard Wordle colors: **green** = correct position, **yellow** = wrong position, **gray** = not in word. The comments highlight strategic thinking about letter frequency, multi-step lookahead reasoning, and appreciation for non-obvious gameplay decisions.

> **Prompt 2:** "You are evaluating the strength of admiration to creativity a comment expresses toward the specific Wordle gameplay being discussed (either the original poster's or the commenter's own play). Rate on a scale from 1 to 10: High admiration (7-10): Strong, clear praise or excitement about the gameplay, even if short (e.g., 'Amazing', 'Nice', 'That's smart'). Low admiration (1-3): No praise, unrelated comment, neutral statement, criticism, or advice. Middle admiration (4-6): Mild or ambiguous engagement—some strategies involved but lack of strong applause."

Both prompts asked for a 1–10 rating; the distribution of the resulting scores is shown in Figure 4. The comments were hand-filtered by the authors to find 173 comments where the authors agreed the comment was expressing that the game was creative in some way. The initial prompts a majority of texts we did not perceive as expressing aesthetic judgment or admiration for creativity, so the manual filtering step was necessary.

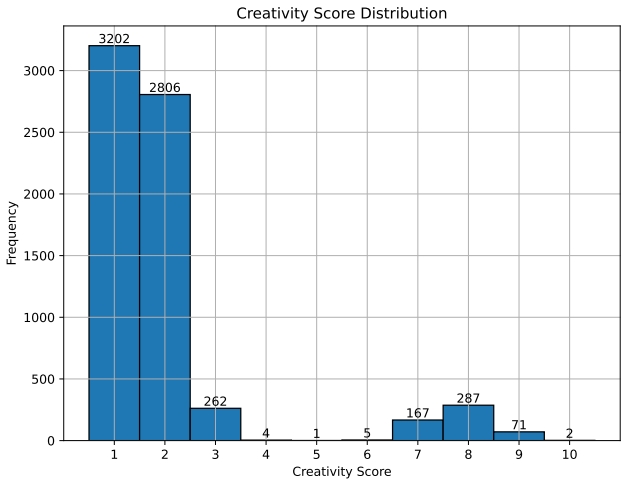


Figure 4: Distribution of creativity scores (1–10) assigned by GPT-3.5 Turbo to Wordle comments. Most comments received low scores (1–3), with a smaller subset receiving high scores (7–9), necessitating manual verification of candidates.

We sample 1,000 games not perceived as creative. Sample comments expressing perception of creativity are shown in Figure 3.

5.2 Methods

We hypothesized that, similarly to chess, perception of creativity in Wordle games is related to the search problem being solved.

The goal in the game of Wordle, for the majority of players, is two-fold: not to "lose" the game by failing to guess the target word in six guesses, and to guess the target word using as few guesses as possible. An optimal strategy that is guaranteed to win while minimizing the number of guesses exists (Bertsimas & Paskov, 2024). A strategy which simply aims to minimize the expected number of remaining possible words after each guess is nearly optimal (Liang et al., 2024).

Define the *current entropy* of a Wordle game as the uncertainty over the remaining possible answers (i.e., $\log N$ where N is the number of possible answers remaining). The goal is to get the entropy to zero (i.e., only one possible answer remains) in as few guesses as possible. This can involve strategic play, where the payoff in entropy reduction is somewhat delayed (in other words, greedy entropy reduction is not always optimal (Bertsimas & Paskov, 2024)).

An operationalization of our hypothesis is that moves that minimize expected entropy over multiple future moves (i.e., "lookahead") would be associated with perception of creativity.

5.3 RESULTS

We hypothesized that the games with evidence of lookahead by two moves would, on average, be perceived as more creative than games without evidence of lookahead.

For each game, we compute the maximum difference/drop between the average entropy one and two moves ahead with optimal play, i.e., -log number of possible guesses left with optimal play after one and two moves, and each current move.

We model the game's being perceived as creative (recall that comments expressing perception of creativity were manually verified from candidates found by our LLMs; games with comments not expressing perception of creativity are randomly sampled) using those two quantities using logistic regression:

$$P(\text{creative} = 1) = \sigma(\beta_0 + \beta_1 \cdot \Delta H_1 + \beta_2 \cdot \Delta H_2) \tag{1}$$

where ΔH_1 is the maximum entropy drop one move ahead with optimal play, and ΔH_2 is the maximum entropy drop two moves ahead with optimal play.

Table 3: Predicting perception of creativity in Wordle.

Variable	**Est.**	**Std.Err.**	***z***	***p***
(Intercept)	-2.537	0.326	-7.780	< 0.001***
ΔH_1	-0.051	0.252	-0.201	0.841
ΔH_2	2.908	1.462	1.988	0.047*

Holding the drop in entropy using 1-step lookahead constant, a drop of 1 bit in entropy using 2-step lookead is associated with an increase in the log-odds of the game's being perceived as creative by 2.908 (SE=1.462, *z*=1.988, *p*=0.047). With a base rate of 50% games being perceived as creative (resampling the negative examples), the model predicts a game receiveing a comment expressing perception of creativity with 83.7% accuracy.

When 2-step lookahead is included in the model, the drop in entropy using 1-step lookahead is not a significant predictor of the game's being perceived as creative (estimate=-0.051, SE=0.252, *z*=-0.201, *p*=0.841).

We find that a drop in entropy when using 2-step lookahead predicts the game's being perceived as creative, whereas a large drop in entropy using 1-step lookahead does not, when controlling for the drop in 2-step lookahead. The maximum drops in 1-step and 2-step entropy are highly correlated (Pearson's $r = 0.8$), so our results indicate that 2-step lookahead is a better predictor of perception of creativity than 1-step lookahead.

We did not run further statistical tests beyond the pre-specified logistic regression, owing to the small sample size.

Our results indicate some perception of creativity in Wordle may be associated with lookahead by two moves, and therefore with search. At 173 positive examples, the sample size is small. Nevertheless, it is interesting that a marginally significant difference was observed. Note that label noise would generally decrease the magnitude of any effect. Future work will build on these preliminary results and will increase the size of the set of Wordle games perceived as creative to increase statistical power.

6 DISCUSSION AND CONCLUSIONS

We have shown indications that perception of creativity is in part to do with search in two different settings. In both, deeper search, and, in chess, more complicated features of the game search tree corresponded to the perception of creativity.

In this work, unlike in previous work, we used a high-quality dataset in chess.

We used a novel dataset of Wordle games. Games which were associated with human perception of creativity were found by filtering comments on Wordle games using GPT-3.5 Turbo and manual verification.

6.1 Operationalizing "Creativity"

In this work, we engaged with human perception of games as "brilliant" in chess and "creative" in Wordle. With chess, we relied on expert annotations, while recognizing that even experts take an "I know it when I see it" approach. In Wordle, we used the authors' own judgment on whether a comment expressed something approximating admiration for creative play (after a filtering pass by GPT3.5 Turbo, resulting in some false negatives).

Operationalizations of concepts such as "brilliance" and "creativity" are inherently imperfect, and, in this work, we did not make a distinction between those related concepts.

6.2 Limitations

The number of positive examples in the Wordle dataset is small (173 games), and in future work we plan to increase this number. Nevertheless, our pre-decided analysis yielded a marginally significant result with a large effect size estimate (with a margin of error).

6.3 Broader Implications

We hypothesize that in some domains of human creativity where search is prominent – games of chess and Go and even Wordle – search — optimization or game tree search — plays an important role in what humans think of as creativity.

Creative domains with formal rules are not limited to chess and Wordle.

Fixed verse is in some ways similar to chess: the game has rules, and one thing that can make the poem be perceived as brilliant is that the author followed the rules while also conveying something to the reader. Just like in chess, you might imagine wondering how a brilliant author came up with the first word in just such a way that a word two stanzas from there fell right into place, and thinking back to how that "plan" could go wrong if a slightly "imprecise" word were chosen along the way.

Prose authors like Georges Perec also used search in the creative process: the novel(s) *La Vie: mode d'emploi* (Perec, 2011), among other things, proceeds to describe rooms in a Parisian apartment block by using the knight's move to skip around the block.

AI use statement

Generative AI was used as a component of our research method: GPT-3.5 Turbo was used to screen `r/Wordle` comments for candidate expressions of perceived creativity, as described in Section 5.1. All candidates surfaced by the model were subsequently hand-verified by the authors, and only hand-verified comments were used as positive examples; the model's scores are not used as labels in any reported analysis. Claude was used to help with writing, editing and coding. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

Our Wordle data consists of publicly posted games and comments scraped from the `r/Wordle` subreddit. We report only aggregate statistics and quote comments without usernames or other identifying information; we do not attempt to identify or profile individual users. The chess data comes from a commercially published, expert-annotated game collection (Šahovski Informator, 1966–).

Reproducibility statement

Section 3.2 specifies the search-tree feature construction, including the engine weights and the five node budgets used, and Section 4 gives the architecture and the hyperparameters selected by the 5-fold cross-validated grid search. Section 5.1 gives the full text of the two prompts used to screen Wordle comments and the hand-verification procedure that produced the 173 positive examples, and

Section 5.2 defines the entropy and lookahead quantities entering the logistic regression reported in Table 3.

REFERENCES

Benton J Anderson and Jesse G Meyer. Finding the optimal human strategy for wordle using maximum correct letter probabilities and reinforcement learning. *arXiv preprint arXiv:2202.00557*, 2022.

Dimitris Bertsimas and Alex Paskov. An exact solution to wordle. *Operations Research*, 2024.

Margaret A. Boden. Creativity and artificial intelligence. *Artificial Intelligence*, 103:347–356, 1998.

Ted Chiang. Why A.I. isn't going to make art. *The New Yorker*, 2024. URL https://www.newyorker.com/culture/the-weekend-essay/why-ai-isnt-going-to-make-art.

Simon Colton, John Charnley, and Alison Pease. Computational creativity theory: The FACE and IDEA descriptive models. In Dan Ventura et al. (eds.), *Proceedings of the International Conference on Computational Creativity*, pp. 90–95, 2011.

Aaron Hertzmann. Can computers create art? In *Arts*, volume 7, pp. 18, 2018.

Azlan Iqbal. Computing the aesthetics of chess. In *AAAI Workshop on Computational Aesthetics*. AAAI Press Boston, MA, 2006.

Erik Jenner, Shreyas Kapur, Vasil Georgiev, Cameron Allen, Scott Emmons, and Stuart J Russell. Evidence of learned look-ahead in a chess-playing neural network. *Advances in Neural Information Processing Systems*, 37:31410–31437, 2024.

Jonathan Levitt and David Friedgood. *Secrets of Spectacular Chess*. Batsford chess library. Henry Holt, New York, 1st american ed. in edition, 1995.

Gary Liang, Adam Kabbara, Cindy Liu, Ronaldo Luo, Kina Kim, and Michael Guerzhoy. Semantic, orthographic, and morphological biases in humans' wordle gameplay. *arXiv preprint arXiv:2411.18634*, 2024.

Lichess. Lichess study: Chess studies and analysis. https://lichess.org/study, 2025. URL https://lichess.org/study. Accessed: 2025-08-14.

Ronaldo Luo, Gary Liang, Cindy Liu, Adam Kabbara, Minahil Bakhtawar, Kina Kim, and Michael Guerzhoy. Automatically detecting amusing games in wordle. In *Proc. of the International Conference on Computational Creativity*, 2025.

S. Margulies. Principles of beauty. *Psychol Rep*, 41(1):3–11, aug 1977.

Georges Perec. *La Vie mode d'emploi*. Fayard, 2011.

Šahovski Informator. *Chess Informant*. Šahovski Informator, Belgrade, Serbia, 1966–. Founded by Aleksandar Matanović and Milivoje Molerović. Published biannually (1966–1990), triannually (1991–2011), and quarterly (2012–present).

Brian Patrick Walls. *Beautiful Mates: Applying Principles of Beauty to Computer Chess Heuristics*. Batsford chess library. Dissertation.com, [United States], 1st american ed. edition, 1997.

Josh Wardle. Wordle. Web, 2021. URL https://www.nytimes.com/games/wordle/index.html. Online game.

Kamron Zaidi and Michael Guerzhoy. Predicting user perception of move brilliance in chess. In *Proc. of the International Conference on Computational Creativity*, 2024.