Semantic, Orthographic, and Morphological Biases in Humans' Wordle Gameplay

Authors

Jiadong (Gary) Liang (University of Toronto; esgary.liang@mail.utoronto.ca)
Adam Kabbara (University of Toronto; adam.kabbara@mail.utoronto.ca)
Cindy Liu (University of Toronto; cindyjy.liu@mail.utoronto.ca)
Ronaldo Luo (University of Toronto; ronaldo.luo@mail.utoronto.ca)
Kina Kim (University of Toronto, IBM; giyeon.kina.kim@ibm.com)
Michael Guerzhoy (University of Toronto; guerzhoy@cs.toronto.edu)
DOI: 10.1145/3774399.3774406

Abstract

We show that human players' game play in the game of Wordle is influenced by the semantics, orthography, and morphology of the player's previous guesses. We demonstrate this influence by comparing actual human players' guesses to near-optimal guesses, showing that human players' guesses are biased to be similar to previous guesses semantically, orthographically, and morphologically.

Introduction

Wordle is a daily word-guessing game where players attempt to identify a hidden five-letter word within six attempts (Wardle, 2021). Players usually attempt to minimize the number of guesses they make. Players also usually want to maintain a "streak" of having solved the game within at most 6 guesses for several days.

We explore the difference between near optimal play and human gameplay, which may be influenced by cognitive shortcuts and bi ases. In order to estimate near-optimal plays, we use the maximum-entropy heuristic. We verify that the heuristic is near-optimal. In settings where word association is important, humans are known to be influenced by salient past information, a phenomenon known as *priming* in psychology (Schacter & Buckner, 1998). We conjecture that priming effects exist in the game of Wordle as well. Additionally, we conjecture that humans will tend to depart less from previous guesses in order to minimize cognitive load.

We review the prior work on priming in psychology, and in particular on how priming influences future word choice. We then review the optimal strategy in Wordle, as well as heuristics that approximate it. We introduce our human guess data. We then present our approach to measuring human biases in Wordle gameplay and demonstrate the systematic differences between human plays and near optimal play.

Background: Human Cognitive Processes

Priming is a phenomenon in psychology where past experience influences behavior without the person's explicit knowledge of the influence (Schacter & Buckner, 1998). Specifically, one aspect of priming is word association. Prior works have demonstrated that the grammatical class, semantic

meaning and rhyme of the previous (cue) word would influence the later (response) word by humans.

Deese (1962) conducted early research on word association, exploring the influence of the grammatical class of cue words over word association on the next word. De Deyne and Storms (2008) followed up the study and suggested that no matter whether a noun, a verb, or an adjective are given as cues, the resulting association is most likely to be a noun. Furthermore, for noun cues, while still being dominant, the effect of paradigmatic association (associating with the same class of noun) would decrease when changing from first to second and third responses.

Steyvers and Tenenbaum (2005) demonstrate that an undirected free association network — constructed from data by D. Nelson (1999) that collects human participants' first responses associated with given cue words — where each word is a node and two words are connected if there exists a cue-response pair consisting of those two words — reveals that, on average, each word is connected to only 0.44% of the overall dataset. This finding underscores the sparseness of the association network where the probability of each word being the response given a cue word is not equally distributed.

Steyvers and Tenenbaum (2005) also use data collected by Miller (1995) and Fellbaum (1998), and found that the word network constructed based on semantics of words ex hibits sparseness, connectedness, neighbor ing clustering and power-law degree distribu tion, which are the same characteristics exhibited in the free association network, just to a varying degree.

Bullinaria and Levy (2007) and McDonald and Lowe (2022) observe the connection between information regarding lexical semantics and patterns of word co-occurrence. De Deyne and Storms (2008) also illustrate that the basic semantic features (coded in Wu and Barsa lou (2009)): "taxonomic," "entity," and "situation" are influential in terms of association responses, with "situation" being the most prominent.

D. L. Nelson et al. (1987) demonstrated the effect of rhyme on memory and word association. They run an experiment where subjects would initially study (read aloud) the cue target pair of a given rhyme; then 1.5-2 minutes after they finished studying, a meaning related cue word and its semantic relation with the target word would be given and the participants would be required to read it aloud and recall the word they studied (D. L. Nelson et al., 1987). In the experiment, cue words that rhyme with many other words would decrease the accuracy of the respondent, regardless of the meaning-related cue word (D. L. Nelson et al., 1987). Through conducting a further experiment that changed all the cue-target pairs studied to be meaning-related and only half to be also rhyme-related, D. L. Nelson et al. (1987) showed that the effect of rhyming appears only if the subjects actively attend to it when studying the word pairs.

Matusevych and Stevenson (2018) studied human word association based on word attributes.

Background: Wordle solving mechanisms

The objective of Wordle depends on the player — it can be maintaining the streak (i.e. try not to lose today's game), winning in as few guesses as possible, or even winning the game using funny words.

Most of the solving mechanisms are designed to optimize objectives regarding the number of guesses, such as minimizing the average number of guesses, minimizing the number of guesses in the worst case, etc. Those mechanisms can be classified into two classes: the exact optimization approach and heuristic approaches. The best approaches based on heuristics achieve results that are only marginally inferior to exact methods.

Bertsimas and Paskov (2024) found an optimal and efficient solution for Wordle that minimizes the average number of guesses using dynamic programming. Bertsimas and Paskov (2024) show that the word "SALET" is the best starting guess and the minimum

average number of guesses required is 3.421. They demonstrate that under this approach the program never loses (i.e. it always completes the game within 6 guesses).

Heuristic approaches to Wordle do not guarantee an optimal result but are relatively competitive, and can achieve performance that is very close to optimal. Doddle's minimax heuristic aims to minimize the number of guesses for the worst-case scenario with search depth of 1 (for each guess, it is only optimizing over all the situations after that single guess). For each guess, it iterates through all possible words in the game and chooses the one that minimizes the size of maximum partition (the amount of possible solutions after the current guess) as the guess. Given the starting guess as "SALET", it is guaranteed to finish the game in 5 guesses and have the average number of guesses to 3.482 (Cross, 2022). Doddle's entropy-based heuristic (also with depth 1) reduces the uncertainty at each step by choosing the guess that decreases (on average) the most number of potential solutions after that guess (Shannon, 1948) (Cross, 2022). It is also guaranteed to complete the game in 6 guesses and have the average number of guesses of 3.432.

Data

The human guess data was sourced from Reddit. The machine-generated guesses were obtained using Doddle, an open-source Wordle solver introduced earlier. Although an ideal comparison would be with the optimal model, the Doddle solver was chosen for computational reasons. It's important to note that the performance difference between the exact dynamic programming solution and the heuristic entropy solver is minimal: the exact solution achieves a minimum average of 3.421 guesses, while the heuristic-based solver has an average of 3.482 guesses for its minimax heuristic and 3.432 guesses for its entropy-based heuristic. Doddle's heuristic min-entropy solver that we use will be referred to as *near-optimal*.

Data collection

The data for this research project was collected from the r/Wordle subreddit, where people share their guesses online, contributing to a total of 83,000 data entries (Watchful1, 2023). We used regular expressions to extract guesses and colored square results from posts written in the standard Wordle format.

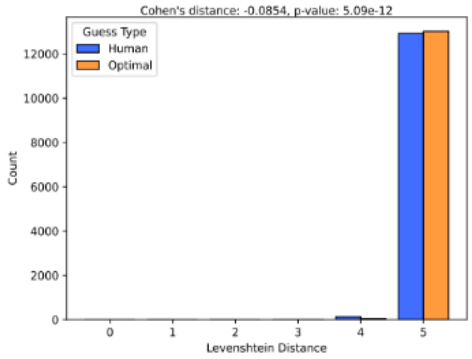
Methods

Measuring Human Biases

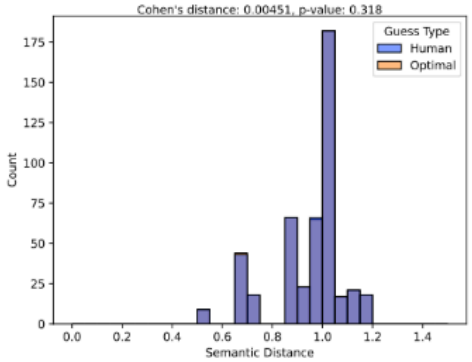
To quantitatively assess the influence of human cognitive biases in Wordle games, human plays are compared to their entropy based near-optimal counterparts, where five different metrics described below are utilized to reveal different aspects of human biases (semantic, orthographic, and morphological). For each guess in the data, the metrics below are computed through comparing that guess with the previous one (instead of comparing with all prior guesses) unless otherwise stated.

Levenshtein Distance The Levenshtein Distance measures the minimum number of edits — insertions, deletions, or substitutions — needed to transform one word into another (Levenshtein, 1966). This feature captures how closely a player's subsequent guesses align with their previous ones in terms of structural similarity. A smaller Levenshtein distance indicates that the player is selecting guesses that are more similar to their prior attempts, potentially reflecting a reluctance to explore novel letter combinations or a preference for minimizing cognitive effort.

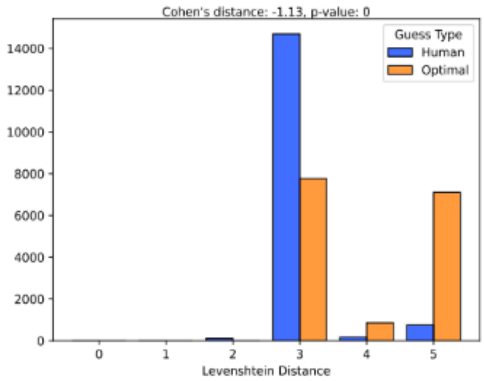
Semantic distance The GLoVe distance is computed using negative cosine similarity between word embedding pairs. Words are represented as vectors using GLoVe, and GloVe distances are computed using negative cosine similarity (Pennington et al., 2014).



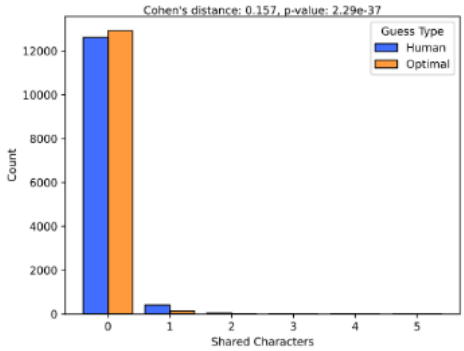
(a) Levenshtein distance between human guesses and near-optimal guesses for 0g0y5b: both choose distance 5 most of the time. Humans subotimally play letters they know aren't there. Note: reference the bottom of this page for Wordle notations.



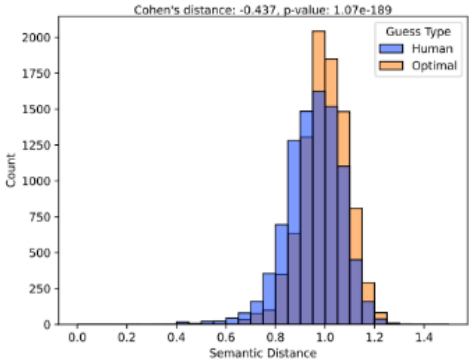
(d) Semantic distance between human guesses and near-optimal guesses for 3g2y0b: no bias.



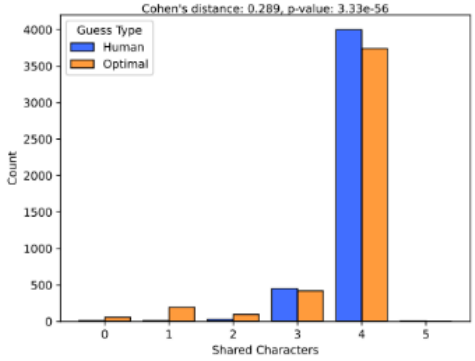
(b) Levenshtein distance between human guesses and near-optimal guesses for 2g0y3b: humans under-explore compared to near-optimal.



(e) Character-level difference between human guesses and near-optimal guesses for 0g0y5b: humans play very obviously suboptimally by reusing characters they know are not there.



(c) Semantic distance between human guesses and near-optimal guesses for 0g0y5b: humans slightly biased towards underexploring.



(f) Character-level difference between human guesses and near-optimal guesses for 2g2y1b: humans play sub-optimally by using new characters.

Figure 1: Notation (c_g, c_y, c_b): c_g - number of "green" guesses (correct letter in the correct place); c_y - number of "yellow" guesses (correct letter in the incorrect place); c_b - number of "black" guesses (incorrect guess).

Character-level difference measures the extent to which players deviate from their initial guesses, quantified by the number of differing characters between subsequent guesses. More character-level difference suggests a greater willingness to explore alternative solutions. Conversely, minimal deviation indicates an over-reliance on early guesses.

Rhyme To determine whether two words rhyme or not, their phonetic transcription was used. This was achieved with the help of the pronouncing library, which provides a phonetic transcription based on the CMU Pronouncing Dictionary "The CMU Pronouncing Dictionary" (2015) . Two words are considered to have a *perfect rhyme* if they have matching phonetic endings which include stressed vowels. We assess if the guess rhymes with the previous one.

Cohen's d Cohen's d is a measure of effect size that quantifies the standardized difference between two means, in this case, human and model performance (Sullivan & Feinn, 2012). Cohen's d transforms the absolute difference between means into standard deviation units, enabling a direct comparison of the magnitude of this difference across various metrics. Effect sizes are traditionally classified as small (d = 0.2), medium (d = 0.5), and large ($d \geq 0.8$) (Carson, 2012)

Experiments

We compare how human guesses/plays differ systematically from near-optimal play. We obtain distributions of human plays and near optimal plays, and compare them. We assess the effect size using Cohen's d, and we computed the p-values based on the t-statistics for the difference between the two distributions.

We analyze separately games starting from different positions. We use the notation $c_g c_y y c_b$, where c_g denotes the number of "green" guesses (correct letter in the correct place), c_y denotes the number of "yellow" guesses (correct letter in the incorrect place), and c_b denotes the number of "black" guesses (letter guesses that are incorrect).

In Figure 1, we present some observations from the results of our comparison of human play with near-optimal play for specific configurations. We observe that in many, though not all cases, humans are biased towards their previous guesses — especially early in the game, when many valid options remain. This suggests that human gameplay, which combines creativity with optimization, tends to reflect more about players' strategic tendencies when the solution space is still large.

We additionally report that the optimal guess rhymes with the previous guess 7.3% of the time, but humans make a guess that rhymes with the previous guess 9.3% of the time (p value < 0.001).

Conclusions

Human gameplay in Wordle exhibits a bias toward previous guesses semantically, orthographically, and morphologically (e.g., sharing the last syllable). The bias is systematic, indicating human gameplay does not merely randomly deviate from optimal play. Initial work indicates that those biases affect human perception of Wordle games (Luo et al., 2025). Our findings can influence game design, especially as it pertains to optimizing the experience of games, as well as understanding how and why people enjoy word games that involve word association.

References

Bertsimas, D., & Paskov, A. (2024). An exact solution to wordle. *Operations Re search*.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior*

research methods, *39*, 510–526.

Carson, C. (2012). The effective use of effect size indices in institutional research. *31st Annual Conference of the North East Association for Institutional Re search*, *41*, 41–48. The CMU pronouncing dictionary [Accessed: October 13, 2024]. (2015).

Cross, A. (2022). Doddle. *https://github.com/CatchemAL/Doddle*.

De Deyne, S., & Storms, G. (2008). Word associations: Network and semantic properties. *Behavior research methods*, *40*(1), 213–231.

Deese, J. (1962). Form class and the determinants of association. *Journal of verbal learning and verbal behavior*, *1*(2), 79– 84.

Fellbaum, C. (1998). Wordnet: An electronic lexical database. *MIT Press*, *2*, 678– 686.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Proceedings of the So viet physics doklady*.

Luo, R., Liang, G., Liu, C., Kabbara, A., Bakhtawar, M., Kim, K., & Guerzhoy, M. (2025). Automatically detecting amusing games in wordle. *Proceedings of the International Conference on Computational Creativity*.

Matusevych, Y., & Stevenson, S. (2018). Analyzing and modeling free word as sociations. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *40*.

McDonald, S., & Lowe, W. (2022). Modelling functional priming and the associative boost. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, 675–680.

Miller, G. (1995). Wordnet: An on-line lexical database [special issue]. *International Journal of Lexicography*, *3*(4).

Nelson, D. (1999). The university of south florida word association norms. *http://w3. usf. edu/FreeAssociation*.

Nelson, D. L., Bajo, M. T., & Canas, J. J. (1987). Prior knowledge and memory: The episodic encoding of implicitly activated associates and rhymes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(1), 54.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. http : / / www . aclweb.org/anthology/D14-1162

Schacter, D. L., & Buckner, R. L. (1998). Priming and the brain. *Neuron*, *20*(2), 185– 195.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, *29*(1), 41–78.

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the p value is not enough. *Journal of graduate medical education*, *4*(3), 279–282.

Wardle, J. (2021). Wordle [Online game]. https : / / www . nytimes . com / games / wordle/index.html

Watchful1. (2023). Subreddit comments/submissions 2005-06 to 2023-12. https : / / www . reddit . com / r / pushshift / comments / 1akrhg3 / separate dump files for the top 40k subreddits/

Wu, L.-l., & Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. *Acta psychologica*, *132*(2), 173– 189.

Appendix: Sample Data

This appendix contains a series of visualizations that support the main findings of this study by illustrating key differences between human gameplay and near-optimal model guesses in Wordle. The figures present histograms for several metrics used in our analysis, including Levenshtein distance, semantic distance (Word2Vec and GloVe), shared syllables, shared characters, and rhyme occur rence. Each figure compares human guesses against the optimal model guesses under various game states, such as those with partial or no feedback (e.g., 0g0y5b, 2g0y3b). Metrics such as Cohen's d and p-values are provided to indicate the magnitude and statistical significance of the observed differences.

Ultimately, these visualizations highlight the extent to which human players rely on structural and semantic similarities in their guesses, favoring familiarity over exploration, particularly when faced with partial confirmation of correct letters.

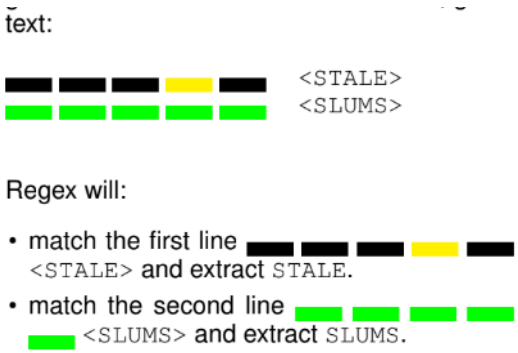
The full data we used is in a further Appendix.

Data cleaning

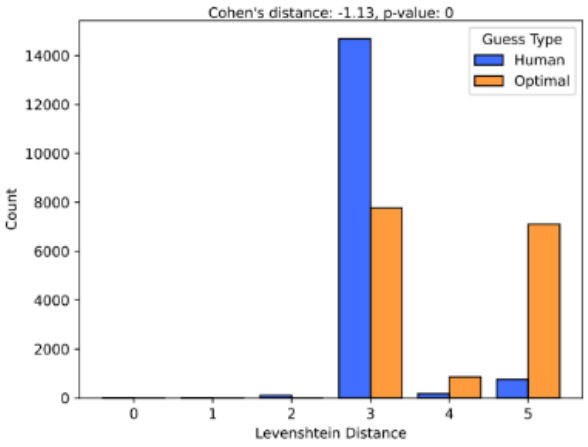
For each guess, the unnecessary parts such as the special symbols (, &!lt;) are removed. To ensure the integrity of the data provided by Reddit users, a cross-referencing process was conducted between the dataset and a Wordle answers database. This approach verified the accuracy of the Wordle IDs submitted and ensured that the answers had not been altered. If a Wordle ID was not provided, the corresponding game was considered illegible, as there was no way to confirm the authenticity of the data. In cases where a Wordle ID was provided without an answer, the last guess was cross-referenced with the Wordle answers dataset. If the last guess matched the correct answer, it was recorded; otherwise, the entry was removed.

To maintain consistency, all guesses were converted to lowercase. The data cleaning process eliminated entries where users did not include their guesses or submitted answers for non-Wordle games. Additionally, any unsolved Wordle games were removed from the dataset. As a result of these cleaning efforts, the dataset was reduced from 83,000 entries to a more manageable 65,000 entries. Ultimately, information about the player, words guessed, and the number of guesses each user made are obtained.

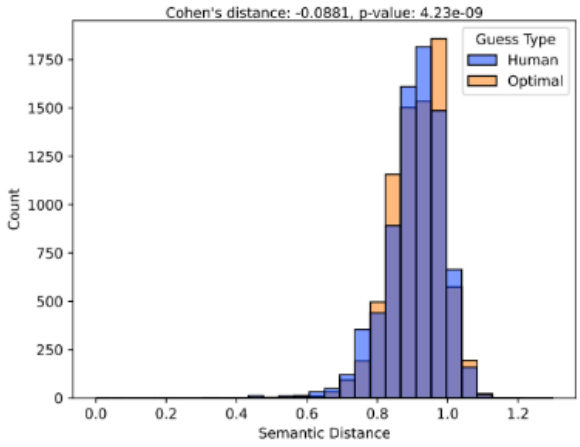
Regex is used to identify lines in Wordle posts where users have displayed both their square results and their guesses. Regex searches for the combination of colored squares and five letter guesses enclosed in special HTML-like tags (<WORD>), ensuring that only complete guess lines are extracted. For instance, given text:



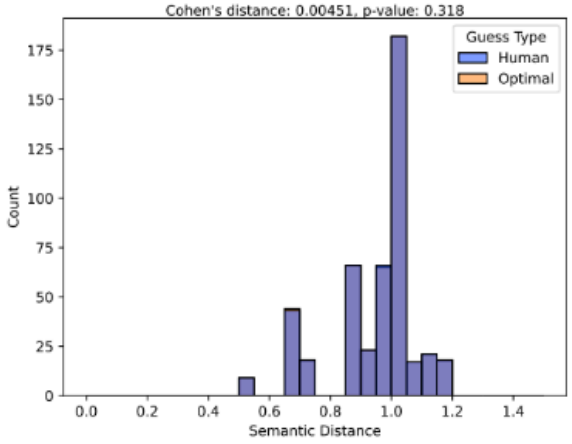
Appendix: more sample histograms



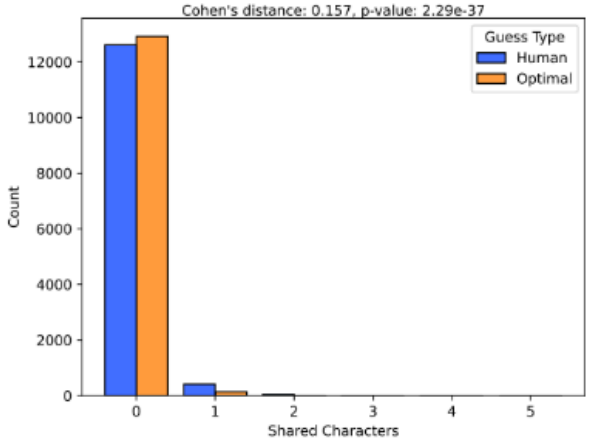
Levenshtein distance histogram for 2g0y3b



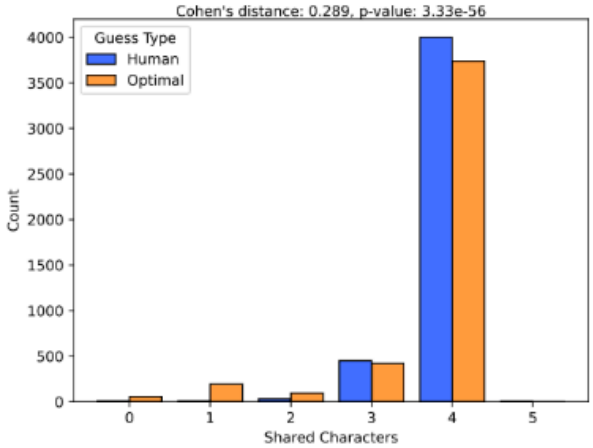
Word2vec distance histogram for 0g0y5b



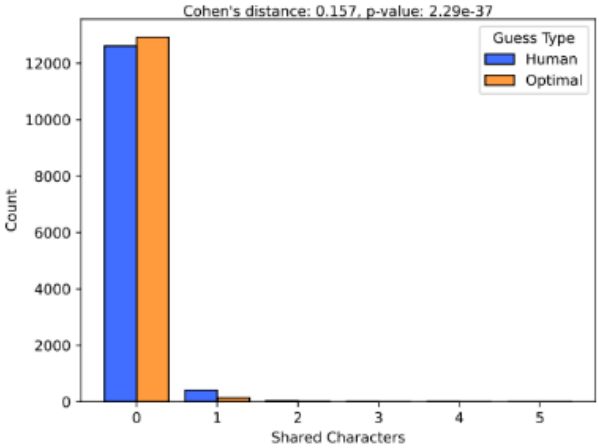
Glove distance histogram for 2g3y0b



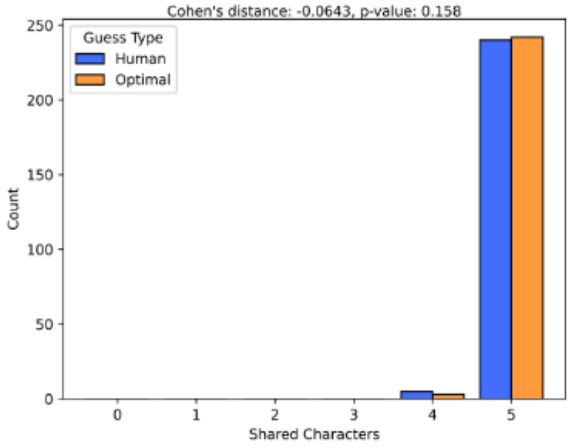
Common syllables histogram for 0g0y5b



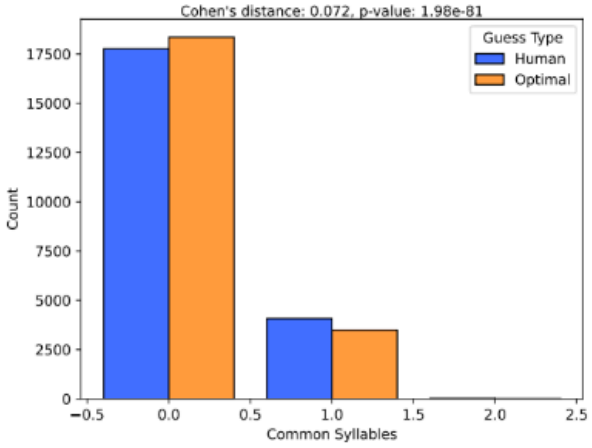
Shared chars histogram for 2g2y1b



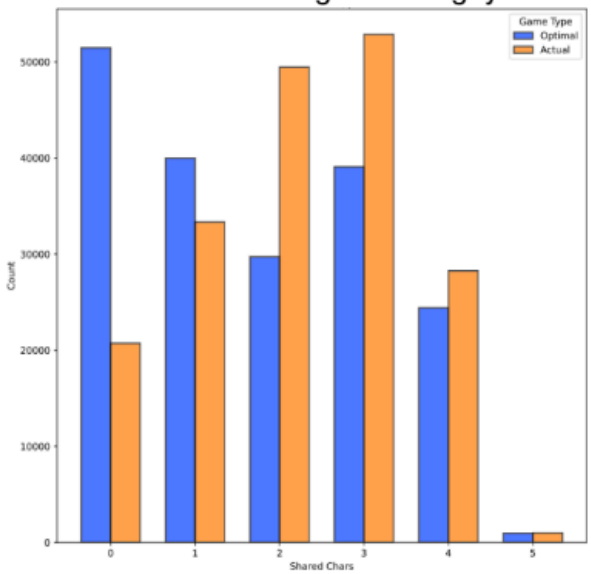
Shared chars histogram for 0g0y5b



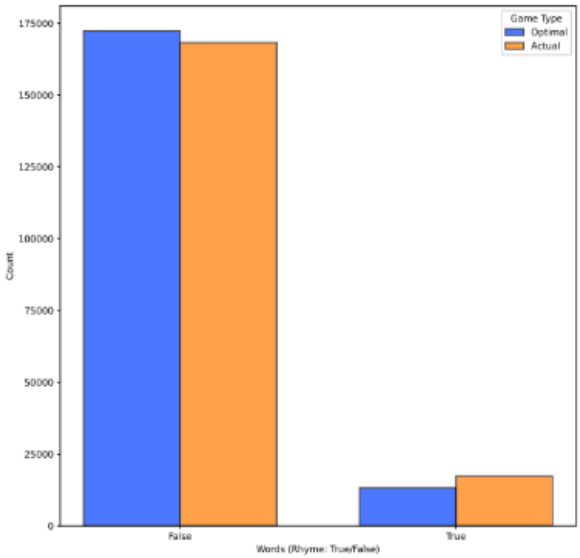
Shared chars histogram for 0g5y0b



Common syllables histogram for 3g0y2b



Total shared chars histogram



Proportion of rhyming guesses

Jiadong (Gary) Liang is a third-year undergraduate student at the University of Toronto, majoring in Machine Intelligence within the Engineering Science program. His research interests lie in machine learning, optimization, and reinforcement learning, with a focus on both theoretical foundations and practical applications of AI.

Adam Kabbara is a second-year engineering Science student majoring in Electrical and Computer Engineering at the University of Toronto. His interests lie in artificial intelligence, embedded systems, and applied machine learning, especially in robotic systems. His work spans research, engineering for social impact, and leadership roles in design teams.

Jiaying (Cindy) Liu is pursuing a Bachelor's degree in the Engineering Science program at the University of Toronto, specializing in Machine Intelligence. Her research interests encompass machine learning, mathematical optimization, and operating systems.

Ronaldo Luo is a third year Engineering Science undergraduate student majoring in machine intelligence at the University of Toronto. He is interested in the general area of machine learning and especially in reinforcement learning and AI safety.

Kina Kim is a Db2 Cloud SWE at IBM and a recent graduate from the University of Toronto in Machine Intelligence. Her undergraduate thesis focused on optimal trajectory planning for unmanned ground vehicles using RL and GNN. Her research interests surround AI applications, using Graph and RL approaches for better human-AI interaction.

Michael Guerzhoy is an Assistant Professor, Teaching Stream in the Division of Engineering Science and the Department of Mechanical & Industrial Engineering at the University of Toronto. His principal interests are in teaching introductory computer science and machine learning, as well as in research in applied ML and neural network interoperability.