Final Project: Evaluating Language Model Performance on NYT Connections

Christopher Bravo
cbravo8@gatech.edu

Abstract

The purpose of this research is to explore how language models perform on the New York Times word puzzle game, Connections. Through this analysis it was determined that training can be utilized to improve language model performance; however, overall, language models naturally struggle with the game due to its usage of ambiguous associations between words. In future research, it may be feasible to test adaptations to language models which will counteract these natural limitations.

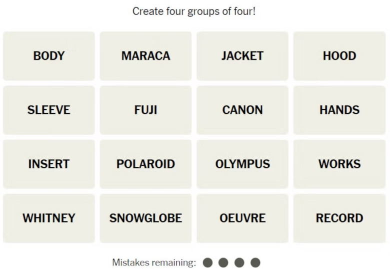
1. Introduction

1.1. Connections

With the growing interest in cognitive health, there has been a corresponding surge in the popularity of cognitive games. Games like Wordle and crossword puzzles present users with complex mental challenges and puzzles that encourage creative thinking. Released by the New York Times in 2023, one such game Connections, works by challenging players to identify the shared connections between groups of words. Each Connections puzzle features 16 words, with each word belonging to one of four categories. On each attempt in the puzzle, users select four words they believe have a connection. They continue until all four connections in the game have been identified.

Connections in the game vary in difficulty to detect. Yellow connections are the most straightforward, often representing examples of a widely known category (e.g. colors, shapes). Blue and green connections are more challenging, including more obscure categories of words in pop culture, history, or entertainment (e.g. Pokemon characters, U.S Presidents last names). Purple connections are by far the most challenging, comprising of plays on words or particularly obscure categories (e.g. words with double letters, countries ending in the letter ”a” etc).

As a verbal-based cognitive task, solving the game Connections can be modeled according to connectionist cognitive frameworks, particularly the Construction-Integration model. (1)

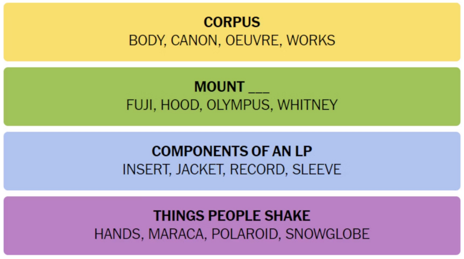


Create four groups of four!

BODY	MARACA	JACKET	HOOD
SLEEVE	FUJI	CANON	HANDS
INSERT	POLAROID	OLYMPUS	WORKS
WHITNEY	SNOWGLOBE	OEUVRE	RECORD

Mistakes remaining: ● ● ● ●

Figure 1. Connections Game



CORPUS
BODY, CANON, OEUVRE, WORKS

MOUNT ___
FUJI, HOOD, OLYMPUS, WHITNEY

COMPONENTS OF AN LP
INSERT, JACKET, RECORD, SLEEVE

THINGS PEOPLE SHAKE
HANDS, MARACA, POLAROID, SNOWGLOBE

Figure 2. Connections Answer

The Construction-Integration model is applied in the following manner:

Construction (1)

1. **Surface Code**: Players start by recognizing and understanding each of the 16 words provided in the game.

2. **Textbase**: Players form initial propositions about each word based on their meanings. For example, the word apple” might generate propositions like fruit,” red,” tree,” etc.

3. **Situation Model**: Players create a mental model of possible categories and relationships among the words by using their background knowledge and making inferences.

Integration (1)

1. **Activation of Propositions**: Players activate relevant properties and associations for each word. For example, the word apple” might activate associations like fruit,” food,” orchard,” etc.

2. **Spreading Activation**: Through a network-like process, related concepts are activated. If apple" activates fruit," it may also activate banana," orange," and pear" as related concepts.

3. **Coherence Formation**: Players strengthen the connections that make sense and form coherent categories while discarding irrelevant or contradictory connections. For example, apple," banana," orange," and pear" would form a coherent category of fruits."

One aspect of the game Connections that is particularly intriguing from the lens of the Construction-Integration model is that connections between words often do not have an explicit semantic relationship. For instance, in one instance of the puzzle, each word was a part of a movie title. Although the words "When", "Harry", "Met", "Sally" have no explicit semantic connection, when formed together they create a distinguishable movie title which does give each word a semantic connection (words that form the movie title *When Harry Meets Sally*). These implicit based connections between words create ambiguity for players. Within the Construction-Integration model they must learn how to activate non-trivial associations between words in order to solve the puzzle.

Under this framework, one interesting question which arises is the extent to which training can strengthen the ability to detect non-trivial associations between words. If players receive feedback on the connections they select, they may learn how to distinguish the ambiguous relationships between words which Connections features.

In practice however, testing this research question with live participants presents multiple challenges. Recruited participants may have varying levels of familiarity with word or puzzle games, including Connections, which will vary baseline performance. In addition, participants almost certainly have varying levels of familiarity with the pop culture or entertainment references which are featured in a majority of Connections puzzles. Recruited participants also pose a logistical challenge, in that they likely require a financial incentive and may need additional time and attention to complete the study.

1.2. Language Models

To address the challenge of recruiting live participants one alternative is to deploy and train language models to play the game Connections. Language models generate and predict plausible text based on text input (3), which in turn makes them capable of mimicking human language. Similar to how the Construction-Integration model models the association between words, language models assign probabilities between words occurring in relation to one another. The language model emits language by emitting the highest probability word in a sequence. With this underlying mechanism, language models can play Connections by assigning probabilities to words appearing in categories together and selecting the sequences of words with the highest probabilities of being together, in a way that is fundamentally similar to the Construction-Integration model. From a training perspective, language models are also advantageous in that their weights are continuously updated with text input. In turn, for playing Connections, language models would not need to be retuned after playing each round of the game. For these reasons, language models can serve as a practical substitute for live participants and may even show capabilities of learning the game and improving with exposure and feedback.

Moreover, by using a language model one avoids some of the limitations of human participants. A language model that has been trained on vast quantities of data will have access to all feasible sayings and pop cultures references. In addition, the model can be deployed at relatively low costs and with significantly less oversight than a human participant.

1.3. Research Question

Given the choice to use language models in this research, the primary research question of this project is to explore whether language models can learn how to detect ambiguous associations between words as seen in the game Connections. This research topic contributes to a growing body of research examining the performance of language models on game tasks, and in particular it addresses language models' ability to identify more obscure semantic relationships between words, an area largely unexplored (2).

From a practical perspective, this research may be used to assess whether language models should be utilized for enhancing natural language processing applications that require understanding nuanced and less obvious connections, such as in particularly niche educational tools, content recommendation systems, or advanced search algorithms. Additionally, insights from this study could inform the development of more sophisticated AI in games and interactive applications, where detecting subtle semantic relationships is crucial for improved user experience and engagement.

2. Experiment Design

2.1. Model Selection

To evaluate the performance of large language models, we selected OpenAI's most advanced model, GPT-4o (5). GPT-4o boasts 175 billion parameters and was pre-trained on the largest assortment of text data among OpenAI's models.

These advantages make GPT-4o particularly well-suited for playing Connections, a game rich in cultural references and idioms. Additionally, GPT-4o is optimized to answer questions faster and is more computationally efficient than its predecessors.

2.2. Puzzle Selection

For evaluating language model performance on Connections, 100 daily puzzles were selected from the period of June 14th, 2023 until September 22nd, 2023. (4)

2.3. Test Procedure

To evaluate the effect of training on GPT-4o's performance, two separate instances of GPT-4o were initialized. Both instances were initially instructed on the basic rules of Connections, with preliminary examples provided, emphasizing that connections between words might feature nontrivial associations.

The first instance, referred to as the untrained model, was tasked with completing 100 Connections puzzles, presented in sets of ten. For each puzzle, the model was instructed to provide the name of each category along with the words it believed belonged in that category, according to the rules of Connections.

The second instance, referred to as the trained model, was also instructed to complete 100 Connections puzzles in sets of ten. However, this model received feedback after each trial, consisting of the correct answers to each Connections puzzle.

2.4. Measures of Performance

Model performance for each puzzle was calculated according to two metrics: the percentage of categories correctly classified and the percentage of total words correctly classified.

2.5. Procedure Analysis

3. Results

3.1. Evaluating Untrained Model Performance

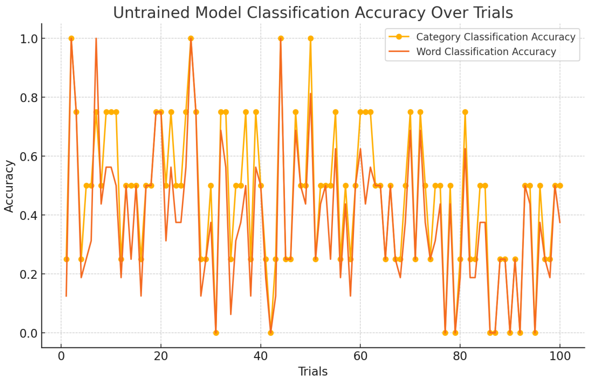


Figure 3. Connections Game

The untrained model demonstrated an average accuracy of 46% (SD = 25%) in identifying connections, while the accuracy for classifying individual words averaged 39% (SD = 24%). Notably, there was a decline in performance over the last 20 trials, with the accuracy for classifying categories dropping to an average of 28% (SD = 22%). This decrease in performance was paralleled by a reduction in word classification accuracy, which averaged 24% (SD = 22%) during the same period.

Puzzles	Accuracy %
Boats, Sandwiches, Cuts of Beef, Nicknames That Are Verbs	0
States of Matter, Edit Menu Commands, Defeat Badly, Anagrams	0
Depart Quickly, Animals that end with X, Shades of Black, Words Before Days of the Week	0
Appetizer Unit, Response to a Correct Answer, Synonyms for Mar, Words which serve as prefixes for Jack	0
Rocky Horror Picture Show, Who Framed Roger Rabbit, When Harry Met Sally, Mad Max Fury Road	0

Table 1. Five lowest word categorization accuracy puzzles for untrained model

Untrained model performance was lowest for connections puzzles which featured categories of connections that involved more obscure usages of the words or implicit relationships the words share (Table 1). Of these puzzles, notably low performing categories included ”Nicknames that are Verbs”, ”Anagrams”, ”Words Before Days of the Week”, ”Words Which Are Prefixes for Jack”, and a puzzle of all movie titles.

In total there were nine puzzles in which the model was unable to classify any words, the majority (7 of 9) which featured connections emphasizing ambiguous associations between words.

Puzzles	Accuracy %
Sneaker Brands, Musicals Beginning with C, Cleaning Verbs, Man Superheroes	100
Sports, Tops, Vegetables, Insects,	100
Seven Dwarfs, File Extensions, Flightless Birds, Tropical Fruits	100
Movie Title Cities, Fashion Magazines, Storms, Cocktails	100
Computer Equipment, Rodents, Musical Instruments, Synonyms For Complain	81.25

Table 2. Five highest word categorization accuracy puzzles for untrained model

Conversely, the untrained model demonstrated high performance for connection puzzles with distinct categories, in particularly scoring perfectly for categories with unambiguous connections (e.g. sports, types of birds, fruits etc).

3.2. Evaluating Trained Model Performance

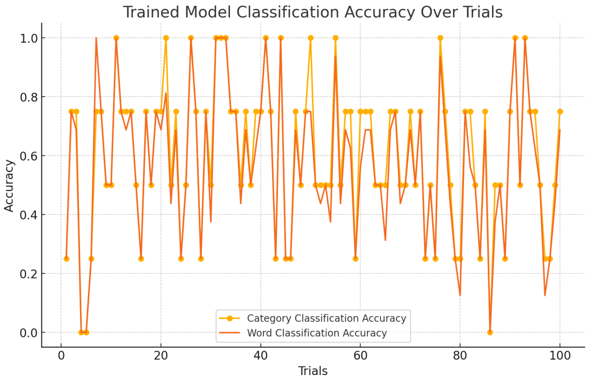


Figure 4. Connections Game

The trained model demonstrated an average accuracy of 60% (SD = 25%) in identifying connections, while the accuracy for classifying individual words averaged 57% (SD = 24%). Notably, the trained model did not demonstrate the same drop off in performance exhibited in the untrained model. In the last 20 trials average category classification accuracy was 55% (SD = 26%) with average word classification accuracy of 50% (SD = 27%)

Puzzles	Accuracy %
Rocky Horror Picture Show, Who Framed Roger Rabbit, When Harry Met Sally, Mad Max Fury Road	0
Monopoly Squares, Shades of Blue, Rappers, Members of a Septet	0
Leg Parts, Baby Animals, Slang For Toilet, Ending in Fish that Are Not Fish	0
Drink Vessels, Woodwinds, American Poets, Consecutive Double Letters	12.5
Intelligent, Airlines, Western Tropes, TV Show Title Surnames	12.5

Table 3. Five lowest word categorization accuracy puzzles for trained model

Trained model performance was lowest for connections with more obscure usages of words, in particular showing difficulty in matching words to movie titles (Table 3). In total, the trained model demonstrated ability to categorize at least a single word in 97 of the 100 puzzles.

Notably, the trained model was capable in certain instances of demonstrating high performance for puzzles with connections designed to bait players. One example of this was in the puzzle with connection groups ”Synonyms For Angry”, ”Things That Are Yellow”, ”Marine Birds”,and ”Words Ending with Boy”. The puzzle featured the word ”Canary” which falls into both ”Things that Are Yellow” and the general category of ”Birds”, however the model was able to register that ”Canary” was not a ”Marine Bird” and made the proper classification into ”Things that are Yellow”.

Puzzles	Accuracy %
Boats, Sandwiches, Cuts of Beef, Nicknames That Are Verbs	100
Synonyms For Angry, Things That Are Yellow, Marine Birds, Words Ending with Boys	100
Hot Drinks, Animal Sounds, Tree Features, Inside Info	100
Airlines, Greek Letters, Silent “G”, Homophones	100
Eye Parts, Counterfeit, Radio Lingo, Songs That are Names	81.25

Table 4. Five highest word categorization accuracy puzzles for trained model

In addition the model demonstrated ability of registering connections which emphasized more obscure features of words. For instance, the puzzle recognized words starting with silent ”G” as being part of the same connection in addition to recognizing homophones.

4. Discussion

4.1. Comparing Model Performance

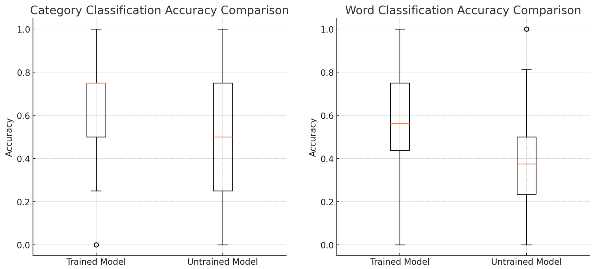


Figure 5. Connections Game

When comparing the results of model performance between the trained and untrained models, it is revealed that some level of training helped the model to identify connections, as the trained model demonstrated significantly higher accuracy in classifying categories (60 % vs 46 %) and words (57 % vs 39 %).

Despite this improved performance compared to the baseline, the trained model did not exhibit evidence of improving as the number of training examples grew, which is somewhat surprising given the nature of learning models to improve with greater training data. One potential explanation for this lack of perceived training improvement is that the puzzles increased in semantic difficulty over time, which effectively neutralized the improved performance of the model. Evidence for this increased difficulty is exhibited in the drop off in performance demonstrated in the untrained model for puzzles 80-100.

4.2. Evaluating Language Model Overall Ability

Although both the trained and untrained language models showed adeptness in identifying simple associations between words, ambiguous connections proved a challenge for both the trained and untrained model (albeit the untrained model struggled more).

The decline in performance for both models on ambiguous connections can potentially be explained by the Construction-Integration model. According to this model, players (in this case the language models) generate propositions for the words in the puzzle. Naturally the instinct is to generate propositions that relate closely to the words themselves. When the language model is confronted with apple, it will naturally create propositions of red or fruit.

However the game Connections challenges the Construction-Integration model in that it groups words by propositions that are seemingly distant from the words themselves. If the word in the puzzle is apple, it is more likely to be related to the technology company (maker of iPhones/Macs etc) or even New York City (i.e. Big Apple) in a Connections puzzle.

For this reason, language models, even with training, struggle playing Connections. Although it is feasible for the models to improve at generating and relating more distant propositions between words, it is not their natural instinct. Instead language models are designed to generate the most likely word to occur next in a sequence.

Another challenge for language models playing Connections is that they rely on word embeddings to generate and understand language. Word embeddings enable language models to group words with similar meanings together through numerical representations (6). Embeddings are determined based on how often words appear around one another. For the game Connections, embeddings prove deceiving for language models for two potential reasons. One reason is that they prevent the model from identifying a word as being a part of a suffix or a prefix, which is a common connection in the game (e.g. the connection ”Words that end in jack” with answers ”pepper”, ”cracker”, ”flapper”, and ”apple”). Since word embeddings are created at the word level, and not the syllable level, any words which share a common prefix or suffix are not inherently categorized as being semantically close to one another and thus are difficult for language models for detect. The second difficult aspect of word embeddings for the game Connections is that often connections are words which rarely appear together or appear in the same context. For instance in the puzzle with movie titles, the words ”When”, ”Harry”, ”Met”, ”Sally” have almost nothing semantically similar to one another, and would only appear together in the context of movie titles (which is a miniscule subset of the training data). In turn language models struggle for these highly-specific examples of words being together.

5. Conclusion

The aim of this research was to explore whether language models can learn to detect ambiguous associations between words as seen in the game Connections. The project findings indicate that while some level of training enhances a model’s ability to identify connections, significant challenges remain, particularly with ambiguous associations.

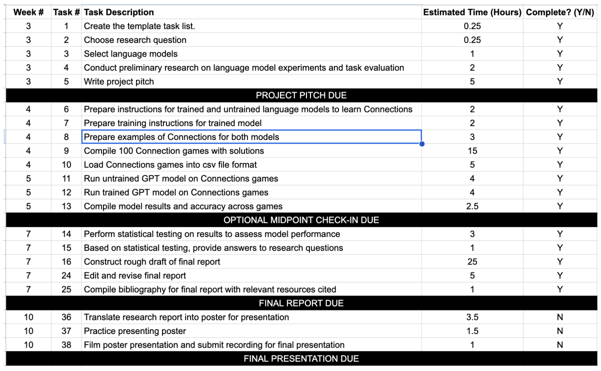
In future research, one potential methodology to improve language model performance on the game Connections is to create alternative embeddings for the language model to utilize. Potential alternative embeddings may include syllable based embeddings so that syllables or groups of syllables used in familiar contexts are represented by the model as being similar to one another. Alternatively, one may consider creating an embedding system based on how words appear together in the game Connections. These artificial enhancements, although somewhat detracting for normal usage of a language model, may improve performance on the game Connections by encouraging language models to explore more distant relationships between words. Moreover, they may be used by the New York Times or other puzzle makers to generate other word puzzle games.

In addition for further research there may be opportunity to understand how human performance on Connections compares to that of language models. Results may reveal a human tendency to connect ambiguous associations between words in manners that language models are not naturally adept at.

6. Limitations

One limitation of this study is that there was insufficient time to replicate results or conduct multiple rounds of testing for GPT-4o models. More rigorous results may be obtained by averaging accuracy measurements over multiple GPT-4o trained and untrained model results’. There was also insufficient time to test other language models performance to determine if these results generalize to other language models.

In addition, future versions of this study may be improved by incorporating more adaptive feedback to the language model, similar to how the game Connections is played, in turn creating a more accurate representation of how language models would play the game.



Week #	Task #	Task Description	Estimated Time (Hours)	Complete? (Y/N)
3	1	Create the template task list.	0.25	Y
3	2	Choose research question	0.25	Y
3	3	Select language models	1	Y
3	4	Conduct preliminary research on language model experiments and task evaluation	2	Y
3	5	Write project pitch	5	Y
PROJECT PITCH DUE				
4	6	Prepare instructions for trained and untrained language models to learn Connections	2	Y
4	7	Prepare training instructions for trained model	2	Y
4	8	Prepare examples of Connections for both models	3	Y
4	9	Compile 100 Connection games with solutions	15	Y
4	10	Load Connections games into csv file format	5	Y
5	11	Run untrained GPT model on Connections games	4	Y
5	12	Run trained GPT model on Connections games	4	Y
5	13	Compile model results and accuracy across games	2.5	Y
OPTIONAL MIDPOINT CHECK-IN DUE				
7	14	Perform statistical testing on results to assess model performance	3	Y
7	15	Based on statistical testing, provide answers to research questions	1	Y
7	16	Construct rough draft of final report	25	Y
7	24	Edit and revise final report	5	Y
7	25	Compile bibliography for final report with relevant resources cited	1	Y
FINAL REPORT DUE				
10	36	Translate research report into poster for presentation	3.5	N
10	37	Practice presenting poster	1.5	N
10	38	Film poster presentation and submit recording for final presentation	1	N
FINAL PRESENTATION DUE				

7. Task Tracker

8. References

References

[1] Walter Kintsch Cathleen Wharton. An overview of the construction-integration model: A theory of comprehension as a foundation for a new cognitive architecture. hhttps://dl.acm.org/doi/pdf/10.1145/122344.122379/, 1991. 1

[2] Gallota et al. Alarge language models and games: A survey and roadmap. https://arxiv.org/html/2402.18659v1, 2024. 2

[3] Google. Introduction to large language models. https://developers.google.com/machine-learning/resources/intro-llms/, 2023. 2

[4] NYT. Connections. https://www.nytimes.com/games/connections, 2024. 3

[5] OpenAI. Chatgpt, 2024. 2

[6] Turing. A guide on word embeddings in nlp. https://www.turing.com/kb/guide-on-word-embeddings-in-nlp/, 2024. 5