Evaluating GPT Models in Categorization Tasks

Elsie Wang
Halıcıoğlu Data Science Institute
University of California, San Diego
e2wang@ucsd.edu

Abstract

Recent advancements in Large Language Models (LLMs), particularly GPT (Generative Pre-trained Transformer), have raised questions regarding their ability to mimic human cognitive processes such as reasoning and categorization. This paper explores categorization, the mental process of sorting objects or ideas into categories or groups based on shared attributes, and evaluates the categorization capabilities of GPT models, specifically GPT-2 and GPT-3.5, using the New York Times Connections games as a benchmark. These games, which involve grouping words based on shared attributes, offer a diverse dataset to assess the models' performance. Results indicate that GPT-3.5 exhibits superior performance, particularly in simpler categorization tasks, but struggles with more complex categories. GPT-2 achieved a mere 0.36% accuracy in correctly categorizing the Connections categories, while GPT-3.5 showed a notable improvement with an accuracy of 37.04%. Overall, this study contributes to understanding the evolving capabilities of LLMs in emulating human-like categorization processes and sheds light on the underlying mechanisms of categorization. The source code can be accessed here: https://github.com/e2wang/gpt-connections.

1 Introduction

In recent years, Large Language Models (LLMs) have witnessed remarkable performance, with advancements in models such as GPT (Generative Pre-trained Transformer) gaining popularity for its linguistic capabilities. Yet, with the rise of LLMs, many wonder how closely these models are able to mimic the human cognitive ability to reason, problem solve, perceive, and learn. This paper attempts to answer this question by assessing LLMs on categorization, a fundamental cognitive process in human understanding and reasoning that involves discerning relationships and grouping items based on shared characteristics or properties. In this study, I aim to evaluate the performance of GPT models, particularly GPT-2 and GPT-3.5, in categorization tasks. By utilizing past New York Times Connections games, which inherently test the ability to categorize words based on shared attributes, I aim to explore how effectively GPT can emulate human-like categorization processes when presented with diverse linguistic stimuli. I hypothesize that GPT-3.5, with its larger size, improved architecture, and extensive training data, will outperform GPT-2 in categorization tasks, particularly in accurately identifying the word that shares the most commonality with a given set of words from the New York Times Connections games. Additionally, I expect GPT-3.5 to exhibit higher probabilities and lower surprisal values for its responses compared to GPT-2, indicating a greater level of confidence and predictability in its categorization judgments.

1.1 Categorization

Categorization is the cognitive process through which individuals organize the world by grouping objects, events, or ideas based on shared features or properties. The paper *Categorization* argues that categorization is fundamental to human cognition, particularly in understanding, learning, inference,

explanation, conceptual combination, planning, and communication [Medin and Heit, 1999]. For instance, categorization can bring old to new in that one may bring relevant knowledge in the service of understanding, or perhaps update or modify the knowledge used. Categorization can also be used as a tool for inference and reasoning in that categorizing some entity allows for predictions concerning its behavior. For example, seeing a young man cleaning the sidewalk with a toothbrush and categorizing him as a “fraternity pledge" would provide some reason to the strange behavior.

However, the way a human categorizes an object or objects is not completely clear. Psychologists generally have three ways of thinking about categorization: "exemplar theory" which suggests that individuals categorize objects by comparing them to specific examples stored in memory, "prototype theory" which is determined by the similarity to a single summary representation for each category, and "rule-based theory" which suggests that individuals categorize objects based on explicit rules or criteria [Hu, 2023]. Yet, categorization is more complex and dynamic and these three theories cannot fully explain how humans categorize. More ambitious research is required in the field of categorization. Evaluating LLMs' categorization performance can not only give insight into the capabilities of LLMs today, but also shed light on the underlying mechanisms of human categorization.

1.2 Related Work

Many studies have evaluated whether LLMs are capable of human-like cognition such as reasoning and understanding. Shapira et al. [2023] evaluated GPT-4 on Neural Theory-of-Mind tasks and found that LLMs struggle with adversial examples, specifically on versions of the false belief task where the container is transparent. In another study, McCoy et al. [2023] found that GPT-4 performs substantially in decoding a simple cipher but expects that LLMs such as GPT-4 will struggle with under some conditions, including reversing strings, ciphers, and counting letters. Rozner et al. [2021] found that LLMs struggle with complex wordplay like cryptic crosswords. Although LLMs demonstrate great performance in many tasks, LLMs still have a long way to go in mimicking human cognitive abilities.

For this particular task, Zhang et al. [2023] attempted to evaluate LLMs, GPT-3 and BLOOM, on essentialist categorization, which is the tendency of individuals to categorize objects or entities based on their underlying essence rather than solely on superficial characteristics. The study used two essential properties: 1) teleological properties – or what something is for – and 2) what something is made of. Results showed that both LLMs tended to align with human judgements of categorization when teleological information was provided, with teleological properties posing more a significance than appearance in categorization judgements. This is similar to how humans rely on underlying purposes and material composition when categorizing objects. The findings from this study extend on the understanding of capabilities of LLMs and how humans to categorize. This paper attempts to expand on this field by providing a larger and more complex dataset on different LLMs and assess in the broader context of categorization rather than just essentialist categorization.

2 Dataset

2.1 Connections

For my dataset, I used past New York Times (NYT) Connections games to assist in categorization [The New York Times]. Connections, released in June of 2023, is a daily game that tasks players to group 16 words into 4 categories with 4 words in each category. Each puzzle has exactly one solution and is meant to be tricky by having words that could fit into multiple categories. All the words in one category share something in common. For instance, 'BLACK,' 'BLUE,' 'YELLOW,' and 'RED' would be in the same group because they are all colors. The game allows for four "mistakes" to group all the categories correctly with guessing each group/category differing in difficulty level. Each group is assigned a color (Yellow, Green, Blue, or Purple), with Yellow being the easiest category and Purple being the trickiest. Categories would not be as simple as having the same amount of letters or all being verbs. Figure 1 shows an example of the game.

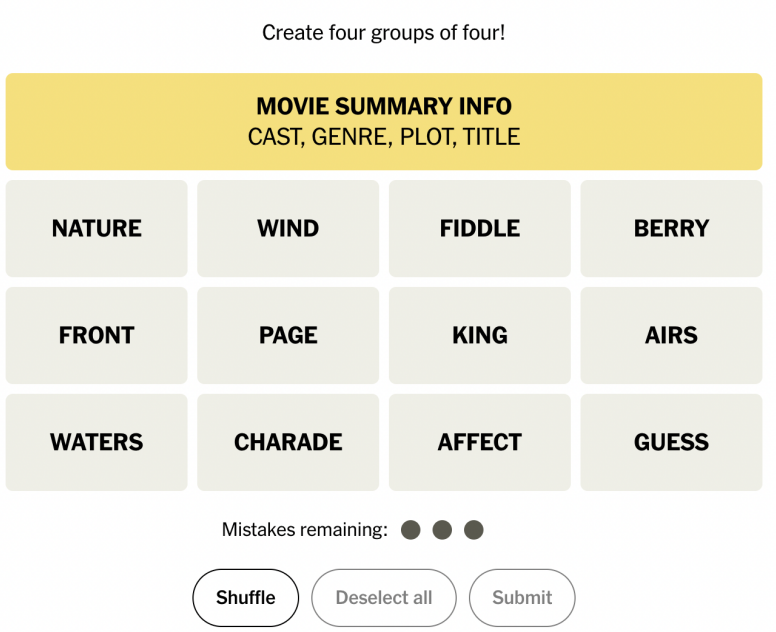

Figure 1: An example of Connections game play. 'CAST,' 'GENRE,' 'PLOT,' and 'TITLE' were guessed correctly to be in the same group, whose category was revealed to be Movie Summary info. Given the yellow background of the category, this category was the easiest to guess. The other three grouped have not been guessed correctly yet.

Categories can vary drastically, even by color. Yellow categories are much simpler and intuitive in that they commonly group words by similarity in meaning or type. For instance, the category 'GENTLE' would group [LIGHT, MELLOW, MILD, SOFT] together due to all the words meaning gentle in some aspect while the category 'ICE CREAM TREATS' would group [FLOAT, SHAKE, SPLIT, SUNDAE] together because all those words are a type of ice cream treat. Yellow categories test for knowledge that most of the general public would be able to solve. On the other hand, purple categories require more critical thinking and creativity, with some categories involving fill in the blank, word structure, grammar, less commonly used or known synonyms or types, or obscure topics. For instance, the category 'SILVER ___' would group [FOX, LINING, SCREEN, SPOON] together while 'THINGS WITH LINKS' would group [CHAIN, GOLF COURSE, SAUSAGE, WEBSITE]. The first category would require associating words with common phrases or descriptions while the second category would require thinking of different definitions of "links."

This game is suitable for categorization tasks due to its diverse array of categories that tests general knowledge. Additionally, with different difficultly levels, the dataset allows us to see how well GPT can categorize, for example, whether it can only categorize yellow (easiest) categories correctly or all including purple categories (trickiest). Since GPT-2 and GPT-3.5 were released before Connections, introducing a dataset that it hasn't been trained on can be an effective in evaluating performance.

2.2 Data Collection

Given that NYT does not release archives of past Connection games, I used Connections+, a third-party site that stores all the Connections games since the beginning of its release [Connections+, 2024]. To collect the answers, I used Selenium Webdriver, a web-based automation framework that allows me to automatically and efficiently play the games on a browser and collect the answers at the end without having to play it myself [Selenium, 2023]. For each game, the driver randomly guessed a group of four words until it ran out of lives and lost, revealing the answers to that game. I collected the answers and their categories from NYT's release on June 12, 2023 up until March 19, 2024, giving me a total of 282 games and 1123 groups of four words, including their difficulty level and category.

3 Methods

For this paper, I used two GPT models: GPT-2 and GPT-3.5, with one beig more advanced than the other. Although they are both are based on the Transformers architecture, GPT-3.5 has a significant edge over GPT-2 in terms of size, performance, and capabilities, making it one of the most advanced and versatile language models available to date. Evaluating both could provide insight into how OpenAI's GPT models, which prove a significance force in the field of LLMs, have developed in categorization tasks.

Models were evaluated on their ability to categorize, more specifically to find the word that has the most common with a given set of words. This was done for each of the 1123 groups of four words, where for the prompt, one word was removed from the group for GPT to solve. Both models were prompted with

```
f"Find␣the␣word␣that␣shares␣the␣most␣in␣common␣with␣{prompts}␣
    among␣the␣following␣words␣and␣ONLY␣the␣following␣words:\n{
    options}.\nGive␣a␣one␣word␣answer.␣E.g.␣'BLUE'"
```

where prompts represented three words belonging to that category and options were the other 13 words used in that Connections game. For instance, if the category was wet weather and the words were ['HAIL', 'RAIN', 'SLEET', 'SNOW'], ['HAIL', 'RAIN', 'SLEET'] would be used as the prompt while 'SNOW' was thrown in the options.

The metrics used to evaluate these models and their responses were accuracy which was whether the GPT model was able to find the word belonging to the category (e.g. 'SNOW'), probability which was the probability of guessing the correct answer or response, and surprisal which was the measure of how unexpected or surprising the response or correct answer was.

3.1 GPT-2

To evaluate GPT-2, I used Hugging Face's Python transformers library with default parameters [Hugging Face]. From preliminary tests, the GPT-2 responses with the highest probability were "mean", "is" and "" which were not included in the options for any of the Connections game. This may be due to GPT-2's inability to process complex prompts. To account for this, the probabilities of all options were calculated and the response with the highest probability was recorded as the model's response. Additionally, the probability and surprisal of the correct answer was recorded.

3.2 GPT-3.5

To evaluate GPT-3.5, I used OpenAI's Python library to get responses from GPT-3.5 [OpenAI, 2024]. For parameters, I used a temperature of 0.7 and maximum tokens of 200. From preliminary observations, GPT-3.5 was able to respond with one-word answers from the list of options or the prompt. However, GPT-3.5 does not provide all the logits like GPT-2, only that of their response. Thus, only the surprisal and probability of the response was recorded.

4 Results

4.1 GPT-2

GPT-2 had only a 0.36% accuracy in correctly categorizing the Connections categories, with the yellow categories (easiest) having the highest accuracy of 0.71% and purple categories (trickiest) having the lowest accuracy of 0.0%. GPT-2 was able to correctly categorize synonyms for "eat," dog breeds, and parts of a river. The yellow categories GPT-2 got correct were two, which both involved synonyms for eating. Probabilities in responses for all color categories were small while probabilities in correct answers were even smaller, close to 0, with marginal difference among categories. Yellow categories had the highest average surprisal of 22.73 while green categories (second easiest) had the lowest surprisal of 22.30. Mean values for all color categories can be seen in Figure 2. The surprisal of responses being correct versus incorrect show marginal difference with responses that were incorrect having a higher surprisal of 18.35 compared to 17.94 with correct responses, as seen in Figure 3.

Color	Response_Probability	Response_Surprisal	Probability_Correct	Surprisal_Correct	Correct
blue	0.000005	18.323693	8.710674e-07	22.393307	0.003546
green	0.000004	18.340462	9.619410e-07	22.298223	0.003546
purple	0.000005	18.399726	7.938414e-07	22.369003	0.000000
yellow	0.000005	18.343368	6.162374e-07	22.734313	0.007092

Figure 2: Mean results for response in GPT-2. For each color category, the table shows the averages in GPT-2 response probabilities and surprisal, the probabilities and surprisal of getting the correct answer, and the accuracy. The colors, Yellow, Green, Blue, Purple, represent the difficulty level in correctly categorizing the group of words in order of easiest to trickiest, respectively.

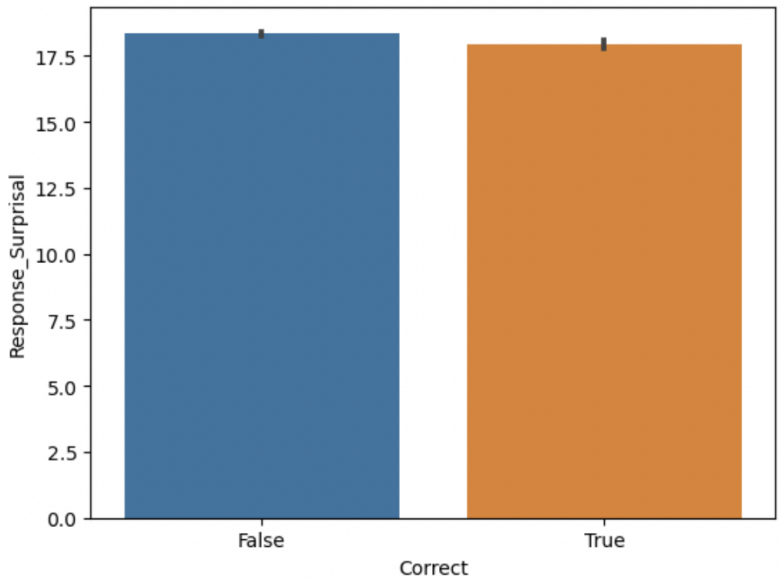

Figure 3: Bar plot of surprisals in GPT-2 responses by correctness. Responses that incorrectly categorized a group of words had a higher surprisal on average compared to responses that correctly categorized a group of words.

4.2 GPT-3.5

GPT-3.5 performed better with an accuracy of 37.04%. Yellow categories (easiest) performed the best with a 62.77% accuracy while purple categories (trickiest) performed the worst with a 8.66% accuracy. Performances improved the easier the category was. By extension, yellow categories had the lowest surprisal of 0.50 while purple categories had the highest surprisal of 1.00. More details can be seen in Figure 4 and visualized in Figure 5. The surprisal of responses being correct versus incorrect show substantial difference with responses that were incorrect having a higher surprisal of 1.06 compared to 0.22 with correct responses, as seen in Figure 6.

Color	Probabilities	Correct	Surprisal
blue	0.716556	0.283688	0.885448
green	0.759423	0.478723	0.624655
purple	0.634872	0.086643	1.001186
yellow	0.823912	0.627660	0.496609

Figure 4: Mean results for response in GPT-3.5. For each color category, the table shows the averages in GPT-3.5 response probabilities and surprisal, as well as the accuracy. The colors, Yellow, Green, Blue, Purple, represent the difficulty level in correctly categorizing the group of words in order of easiest to trickiest, respectively.

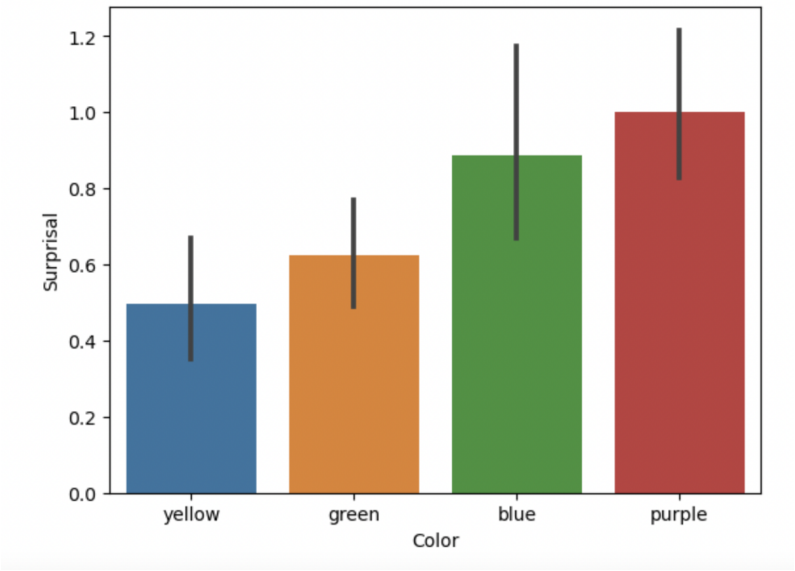

Figure 5: Bar plot of surprisals in GPT-3.5 responses by color. Yellow categories (easiest) had the lowest surprisal while purple categories (trickiest) had the highest surprisal.

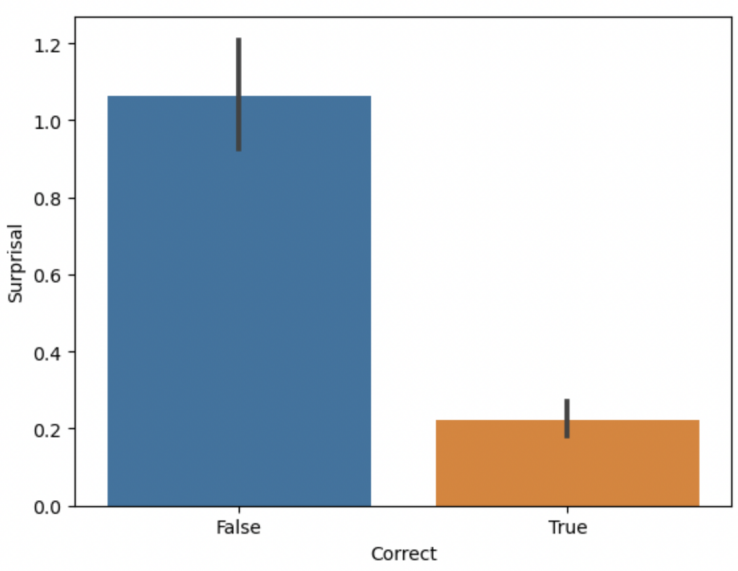

Figure 6: Bar plot of surprisals in GPT-3.5 responses by correctness. Responses that incorrectly categorized a group of words had a higher surprisal on average compared to responses that correctly categorized a group of words.

However, some responses from GPT-3.5 did not fully answer the prompt correctly. Specifically, 16.47% of the responses repeated words from the prompt of three words rather than choosing from the given options. This could be a result of an inefficient prompt, lack of ability from GPT-3.5 to fully comprehend more complex instructions, randomness, or a combination of all. Regardless, filtering out these responses showed similar results at slightly higher accuracies with yellow categories performing the best and purple categories performing the worst.

5 Discussion

As expected, GPT-3.5 outperformed GPT-2 in categorization tasks with a higher accuracy and ability to correctly categorize trickier categories that require more cognition. GPT-2's differences in surprisal by correctness showed minimal difference, perhaps due to its relatively lackluster performance and small sample of correct answers. GPT-3.5 showed lower surprisals and higher probabilities on average with considerable difference in performance by colors. This suggests that GPT-3.5 performs well on categorization tasks for easier categories, but harder categorization tasks still need improvement. With better performance in yellow categories, GPT-3.5 excels in finding simple synonyms or grouping common words by types while it struggles with purple categories, such as filling in the blank to common phrases, association to a word with different definitions, obscure topics, grammatical structure, among other topics that require more cognitive thinking. By correctness, surprisal in GPT-3.5 also showed considerable difference, namely that it assigns lower probabilities when the response is incorrect while it assigns higher probabilities when the response is correct. This may suggest GPT-3.5 has higher confidence in its responses when it is correct than when it incorrect, similar to how humans are more confident in an answer when they know they are correct.

5.1 Limitations

Despite these results, it is important to consider some limitations in this project. First, the Connections dataset is a good starter to assess LLMs in categorization, but it should not be considered a standard benchmark for testing general categorization due to the complexity, different applications, and types of categorization. Additionally, Connections is created by NYT staff which assign difficulties subjectively. This can introduce inconsistencies in difficultly levels and types of categories that may appear. Second, the prompt presented to the GPT models may affect the outcome of the results. Although I tested different prompts to get the desired output structure, there could be parts of the

prompt that may produce more unfavorable or favorable results. The complexity of the prompt could also contribute to the low accuracy for GPT-2 or even GPT-3.5.

5.2 Future Work

One can try prompt engineering to see how different prompts affect the output of the GPT responses before collecting responses. Additionally, a more comprehensive and reliable dataset is required, perhaps to isolate a certain type of categorization or umbrella a larger portion of categorization. With the release of GPT-4, one can also test the latest OpenAI GPT model or other LLMs such as BERT or CLIP. This paper shows that LLMs are increasingly becoming better at categorization, but more research is needed for assessment.

References

Connections+, 2024. URL `https://connectionsplus.io/`.

Mingjia Hu. The cognitive process behind categorizing objects, May 2023. URL `https://blogs.iu.edu/sciu/2023/05/13/process-of-categorizing/`.

Hugging Face. URL `https://huggingface.co/docs/transformers/en/index`.

R Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L Griffiths. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*, 2023.

Douglas L Medin and Evan Heit. Categorization. In *Cognitive science*, pages 99–143. Elsevier, 1999.

OpenAI. Openai/openai-python: The official python library for the openai api, 2024. URL `https://github.com/openai/openai-python`.

Josh Rozner, Christopher Potts, and Kyle Mahowald. Decrypting cryptic crosswords: Semantically complex wordplay puzzles as a target for nlp. *Advances in Neural Information Processing Systems*, 34:11409–11421, 2021.

Selenium, 2023. URL `https://www.selenium.dev/`.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*, 2023.

The New York Times. URL `https://www.nytimes.com/games/connections`.

Siying Zhang, Jingyuan Selena She, Tobias Gerstenberg, and David Rose. You are what you're for: Essentialist categorization in large language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, 2023.