Evaluating the Coherence and Readability of ChatGPT Output Across Multiple Domains

Nick Gable gable105@umn.edu
Stuti Arora arora176@umn.edu
Archit Kalla kalla100@umn.edu
David Burchell burch244@umn.edu

Abstract

ChatGPT is famous for outputting high quality, generally coherent text, but not much research has been done to see the extent to which this is true. Our research tackles this problem, seeking to determine to what extent ChatGPT is coherent and readable when prompted about topics across different domains. We prompted ChatGPT across six different domains, then analyzed coherence / readability using a Neural Coherence Model and the Flesch-Kincaid reading level. We found that ChatGPT was generally coherent across the domains we chose, but that some domains scored lower with occasionally low coherence in prompts across other domains. Our study also highlighted the limitations of coherence modelling across domains with different format or symbol use (in our case, with a lot of algebraic reasoning and symbols).

1 Motivation

As the world becomes increasingly digital, written communication has become an integral part of our daily lives. By analyzing written text and providing feedback on its clarity and organization, readability and coherence models can improve the effectiveness of written communication. The problem under consideration in this project is evaluating the extent to which ChatGPT produces coherent output. While in theory, ChatGPT should produce structured and logically flowing output consistently, ChatGPT may struggle to produce coherent text when prompted with topics it is less familiar with, or with prompts that are less coherent themselves. In our work, we used ChatGPT to generate various texts and measure its output's coherence levels in comparison to standard coherence datasets. From this testing, we want to analyze the ability of ChatGPT to maintain coherence consistently.

Applying readability and coherence models can have significant impacts on the usage and distribution of text in various domains.

- Education: These models can be used to improve how educational materials, like textbooks and instructional videos, are distributed and used by making appropriate materials more accessible to students of different comprehension levels.

- Journalism: Usage of these models can assist in higher quality content with faster turnaround times by Large Language Model (LLM) output from products like ChatGPT is increasingly making an appearance throughout our world. Output from providing automatic feedback on the structure and verbage of text. More so, readability and coherence models can help to reduce implicit biases in writing by identifying potentially problematic language or phrasing.

- Healthcare: Readability and coherence models can help improve patient comprehension of medical information, empowering them to make informed decisions about their health.

- Politics and Government: These models can aid in improving communication with the public by disseminating complex policy information in ways that is easy for the public to understand, which can increase transparency in policy reports or proposals. Additionally, the models can be used to analyze large volumes of public opinion data in order to identify trends and sentiments and further guide policy decisions based on public perceptions.

2 Literature Survey

2.1 Entity-Grid Coherence Models

A 2008 paper by Barzilay and Lapata outlines an entity based approach to modeling text coherence. The paper focuses on *local coherence*, which they define as "text relatedness at the level of

sentence-to-sentence transitions" (Barzilay and Lapata, 2008). Barzilay and Lapata argue that local coherence is important because it is highly tied to the *global (overall) coherence* of a text, something they found to be true in broader research.

Entity grid models work by assuming that "the distribution of entities in locally coherent texts exhibits certain regularities" (Barzilay and Lapata, 2008). By making this assumption (also backed by other linguistic research per the authors), entity-grid models are able to learn about and predict coherent texts without resorting to manual annotation. The model works by forming "entity-grids" that capture the distribution of discourse about entities across sentences. From this grid, entity transition vectors can be extracted, which can then be used to extract individual entities, syntax, and salience, which is then used to generate a score of coherence (Barzilay and Lapata, 2008).

2.2 Neural Coherence Models

Nguyen and Joty in their 2017 paper *A Neural Local Coherence Model* highlight some of the shortcomings of the entity based model and propose a coherence model based on a convolutional neural network that "captures long range entity transitions along with entity-specific features without loosing generalization" (Tien Nguyen and Joty, 2017). Some of the shortcomings of the entity-grid model discussed in the paper include that the model suffers from the "curse of dimensionality" and "feature extraction is decoupled from the target downstream tasks which can limit the model's capacity to learn task-specific features" (Tien Nguyen and Joty, 2017). To overcome this, they propose using Convolution layer in a neural network while keeping a similar feature vector representation from the entity-grid model above. The performance on coherence rating using this method improved the accuracy by up to 3.7. The Hugging Face Coherence (coh) Model we used makes use of "hard negative mining" and extends the Neural Coherence model.

2.3 Readability Models

The earliest version of the Flesch-Kincaid readability formula was to predict the grade of children ranging from third to seventh grade who answered at least 75% correctly about a given passage (Flesch, 1943). It was calculated based on the average number of words per sentence, number of affixes, and number of references to people (McCall, 1925). In the following years, Flesch published the Reading Ease Score (Flesch, 1948) that featured a new formula that replaced previous text measurements with average number of syllables and average sentence lengths. Further adjustment of the weights of the Reading Ease Score by Kindcaid to evaluate readability of technical materials within the Navy, has led to the development of the Flesch-Kincaid Grade Level, which is now heavily used in text simplification evaluation (Kincaid et al., 1975).

2.4 Connection to Our Work

Unlike other work done, we are specifically looking at ChatGPT output. This distinction is important when trying to understand the purpose of our project. Previous research is focused on running compiled datasets through coherence models with the intention of discovering the strengths or weaknesses of that model. In others' work, the coherence model itself is of primary focus, with the content of the data set being secondary. Conversely, our work places far higher emphasis on the data set or input to the coherence model rather than the actual capacities of the model. That is, the reliability of a coherence model is presupposed, and the quality of the data set (ChatGPT responses) is under consideration. Thus, the current practice and existing research has neglected to meaningfully apply coherence models to measure and understand the efficacy of a text generation model like ChatGPT.

3 Approach

Our hypothesis was that ChatGPT may struggle to produce coherent text when prompted with topics it is less familiar with, or with prompts that are less coherent themselves. However, due to the difficulty of coming up with targeted prompts that tested this hypothesis, we moved towards a systematic approach that tested ChatGPT coherence across multiple topic domains with similar prompts.

To start, we picked six distinct topic domains that we would generate prompts for:

- **Operating Systems:** Describing specific operating systems or computer architecture topics, or synthesizing new ones
- **Movies:** Describing movie plots, reviewing them, or creating new ones
- **Cooking:** Describing cooking strategies, or explaining specific processes related to cooking

- **Sports Strategy:** Describing unique strategies in specific sports that build off of existing ones or ideas
- **Fermi Questions:** Open ended mathematics questions with hard to confirm answers
- **Tax Form Explanation:** Explanations of specific tax form needs and details related to them

Following the manual prompt generation of 28 prompts, our process recorded ChatGPT's response to each prompt. We called each prompt separately 6 times using the OpenAI Chat API, without providing context from previous conversations to the model. We also included a few additional test samples from some prompts (around 10 in total) that were generated while testing the code that were averaged into their relevant prompt domains.

Following this, coherence scores (coh) and readability scores were determined by general readability formulas (DiMascio, 2021). The various models were trained and stored using Jupyter Notebooks on the CSE CUDA machines. The primary coherence model used was "Coherence Momentum" from HuggingFace (coh). This model makes use of "hard negative mining" and extends the Neural Coherence model. Measuring readability was done using the Flesch-Kincaid scoring system. This method takes into account the average number of words and syllables of a given text. As a reference, a Flesch-Kincaid score of 9 corresponds to the readability of a vehicle insurance policy.

Analysis and visualization of coherence and readability scores was done using Google Sheets. An LDA model (Gensim)(Prabhakaran, 2022) was used to help summarize data gathered and reveal similarities between coherence results/performance.

Our main challenge in this approach was the prompt generation. Manual prompt generation can be a challenging and time-consuming task. One of the main difficulties is the lack of a clear methodology or set of guidelines for prompt creation, as it often requires domain expertise and creativity. One way we addressed this was to divide domains amongst the group, which would also ensure we would generate unique prompts within topic areas.

4 Experiments

4.1 Key Questions

Our experiments were focused around some key questions: Is there a topic where ChatGPT generate less coherent prompts? Furthermore, are there times where the generated responses were less consistent than others?

4.2 Metrics

4.2.1 Coherence

When evaluating coherence, the output of the models provided scores for each output. An example of some text that would output a low coherence score would be as follows:

> *Insignificance has its world rocked by trees or rocks*

This has a score of 2.44. This sentence is almost entirely incoherent, there is no reasonable way to conceive that insignificance, an idea, can have its 'world rocked', much less by trees or rocks. It is important to distinguish incoherence from nonsensical. A nonsensical sentence still maintains some of the basic properties of a coherent sentence yet would involve things that would be out of place or not reasonable. For example:

> *Should a shark near a sheep grease a robust tense skyscraper?*

This example scores around a 9.58. In this sentence, it is unlikely but still understandable that a shark could be near a sheep to grease a skyscraper. Still though, this scores quite low compared to a coherent sentence which can score in 15+ on the coherence model.

4.2.2 Topic Modeling

Topic modeling is more of a qualitative method of interpreting our results. This would help solving our second key question whether or not the responses generated were consistent within a domain. The goal with this approach was to further understand whether given strictly the generated text, we could consistently identify separate topic domains. To do this we used an Latent Dirichlet allocation (LDA) model. We passed in the generated text as a dataframe and using bi-grams and tri-grams we were able to then use a pre-trained LDA model to analyze the text for topics. In essensce this is a generative statistical model that explains a set of observations through unobserved groups, and each group explains why some parts of the data are similar. The output of this model was a graph where each group represents a set of words in the topic. Topic are extracted by grouping similar words together. A successful result would be that we see

clear clusters in the data for each domain. This also tells us whether or not the responses generated are consistent within a certain domain.

4.2.3 Readability

Readability scores are very basic and do not rely on any machine learning methods. Rather, each of the readability metrics are just hard equations in which our job is only to feed in the parameters. These metrics provides a good understanding on how readable outputs are in comparison to certain domains.

Flesh-Kincaid Grade Level The first metric is the Flesh-Kincaid Grade Level which is defined as:

$$0.39 * \frac{total_{words}}{total_{sentences}} + 11.8 * \frac{total_{syllables}}{total_{words}} - 15.59 \quad (1)$$

To put the result in context, an insurance agreement by law must be at grade level 9 or below according to this test.

Automated Readability Index The Automated Readability Index is another index just to give us a good idea on the age range in which text should be readable by. The equation is defined as:

$$4.71 * \frac{total_{charcters}}{total_{words}} + 0.5 * \frac{total_{words}}{total_{sentences}} - 21.43 \quad (2)$$

The output is a number that corresponds to the appropriate age to read the given text. Anything above 18 is considered a college level.

5 Results and Analysis

5.1 Prompt Analysis

The prompts provided to ChatGPT across the multiple topic areas provide a few interesting insights into the consistency and coherency of the model. OpenAI has a feature on their ChatGPT model to allow for response regeneration to any prompt provided. This functionality was utilized five times on each prompt to give us a total of six responses by ChatGPT to each prompt given. In this way, one can observe the stylistic or coherent consistency of ChatGPT on a single prompt.

As expected, a fair number of the prompts had three responses that received effectively equal coherency scores. However, there were a number of single prompts that received responses that were scored significantly different in coherency. This observation demonstrates the novelty with which ChatGPT is able to pivot in its approach to responding to a given prompt. That is, that a response regeneration request has a substantial effect on how ChatGPT decides to respond to the prompt.

5.2 Coherence Analysis

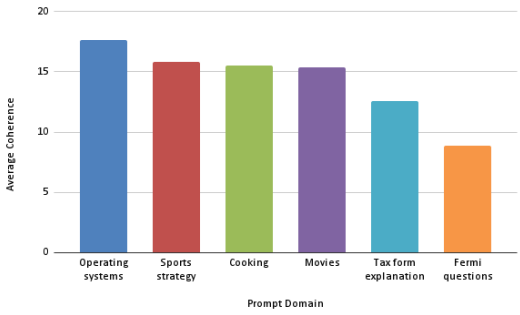

Figure 1: Average Coherence Scores over the Prompt Domains

In Figure 1 we see that the average coherence per domain is fairly similar. We see that operating systems is considerable higher than the other topics. We can attribute this largely to the fact that operating system questions have fairly unique terms that are often seen together. Therefore the overall coherence would be biased higher. The other domains scored similarly.

However the Fermi questions and Tax Forms domain were significantly lower than the others. We can attribute this largely to the fact that the coherence model did not handle numbers and mathematical reasoning well. Fermi questions output were largely in the form of doing calculation and averages. The tax forms responses involved plenty of numbers often relating to salaries or referencing tax codes. We look further into error cases in Section 5.5. Overall, however, we can say that ChatGPT is fairly consistent in generating prompts that are coherent, at the very least against this model.

5.3 Topic Modeling Analysis

In the generated topics we can see some clusters formed around certain topics. For example, in Figure 2 we can see that around the red topic numbered 2 there is a topic numbered 5.

In Figure 3 we see that the most salient terms in topic 2 are definitely related to sports, including terms like player, formation, team, pass, etc. In topic 5, whose word distribution is not shown, the words included are players, hit, shot, ball, etc.

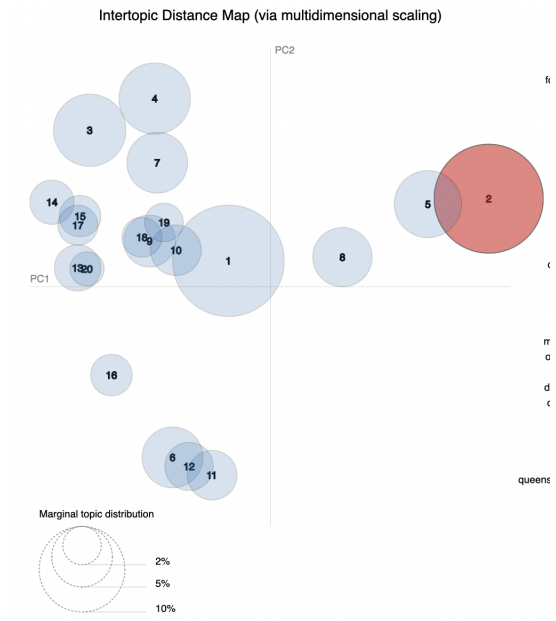

Figure 2: Distance map of different topics, highlighted in red is a topic related to the sports domain

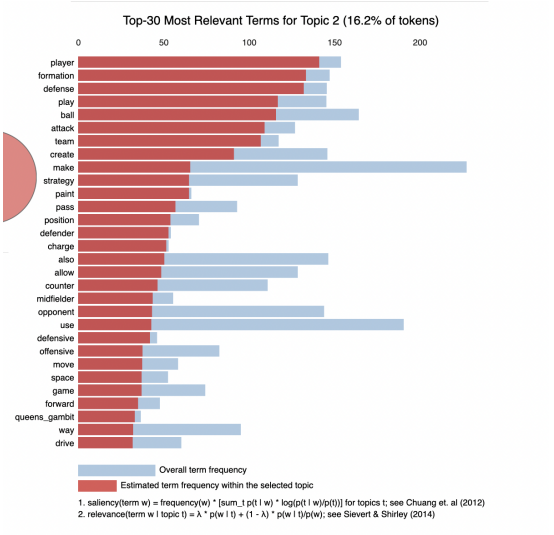

Figure 3: Word distribution for topic number 2, highlighted red in Figure 2

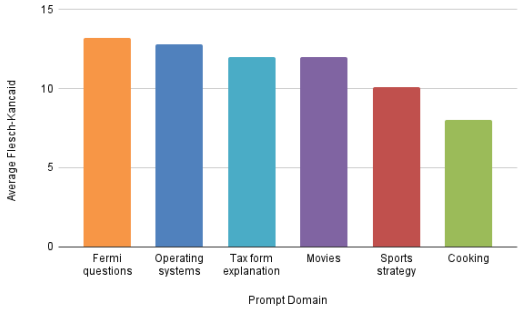

Figure 4: Average Flesch-Kincaid Grade Level over Prompt Domains

With this it follows that they would be close together. Now for topic 8, the words actually happen to be related to football (which is what one of the prompts is about) and thus some of the words like quarterback , receiver, and running back were in those responses, it was classified closer to the other 2 topics, but just not as close. Moving on outside of sports we see that topics numbered 7,3,4 are grouped closely, and that is because they relate to the operating system questions. Topic 14 also relates to operating systems however its word group was more specific towards machine code using terms like getpgrp, error bits, and BSD project. Cooking topics occupy quadrant 3 of the map, and tax was nearly entirely encompassed by topic 1. The remaining however were poorly clustered topics that encompassed the topics movies and Fermi questions. The words in these topics are not consistent. On its own this result for those topics would be concerning, however paired with the coherence metric we find that due to the creative nature of the movie prompts, it follows that perhaps not all words would fall under a consistent topic generated by the LDA. Fermi questions also mixed in with these other topics and largely this is because the context that the questions were provided in were slightly creative in nature, often making up names and items to estimate and such.

5.4 Readability Analysis

The Flesch-Kincaid Grade Level system was applied to all of ChatGPT's responses across the various topic areas. The figure produced illustrates the this Flesch-Kincaid score across each prompt associated within that topic category. With this metric, as opposed to coherency, it is much less dependent on the comparative ratings of other responses to understanding the score of a single response. Overall, the readability results demonstrate that ChatGPT produces more dense responses to more topically difficult prompts (i.e. Operating Systems) when compared to prompts requiring simpler responses (i.e. Cooking).

5.5 Failure cases

One shortcoming in our data is seen in the coherence model's rating of ChatGPT's responses to the Fermi question topic. The average coherence score received is significantly lower than any of the other

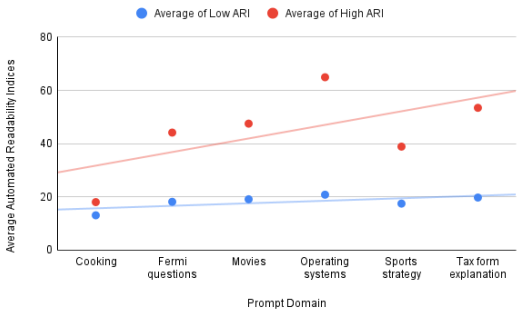

Figure 5: Flesch-Kincaid ARI Average ages over Prompt Domains

topic areas. However, upon human evaluation of these responses, they do not seem that dissimilar or less coherent than ChatGPT's responses to other topics (i.e. Sports Strategy). Therefore, the disparity in coherence score is likely a consequence of the coherence model used. The algebraic and logical content of a reasonable response to a Fermi question is inconsistent with the type of content our coherence model was trained on. The lowest coherence score was for a Math and Logical Reasoning question:

> Helen and Ivan had the same number of coins. Helen had a number of 50-cent coins, and 64 20-cent coins. These coins had a mass of 1.134kg. Ivan had a number of 50-cent coins and 104 20-cent coins. (a) Who has more money in coins and by how much?

A portion of ChatGPT's response was:

> (a) Let's start by finding the value of Helen's coins in dollars: Value of 50-cent coins = 0.5 x number of 50-cent coins Value of 20-cent coins = 0.2 x 64 = 12.8 Total value of coins = Value of 50-cent coins + Value of 20-cent coins = 0.5x + 12.8

It should be noted that ChatGPT actually produced the incorrect answer to this question, although accuracy was not something measured in this project or by the coherence models. This disparity is likely due to the mathematical nature of ChatGPT's responses to Fermi questions. Our coherence model was trained almost exclusively on text and news articles from the Wall Street Journal (coh), so algebraic expressions will inherently seem less coherent to it than any other text passage.

Solving this problem began with its identification via human evaluation of ChatGPT's several responses across all topic areas. Realizing it is a shortcoming of the coherence model used, options of fine tuning or use of an entirely different model are available.

6 Discussion

6.1 Dataset

Our dataset was made in-house, asking generally descriptive or synthesis based prompts related to the domains we chose. As such, we were able to tailor it to suit the specific problem we were trying to solve (coherence / readability).

Our dataset is designed to force ChatGPT to output large amounts of text related to the target domain, with open-ended prompts that would ask questions something along the lines of "describe <topic>" or "write about <subtopic> relating to <topic>". Our prompts also often included some sort of synthesis component that required ChatGPT to talk about something novel, the intent being to prevent it from simply paraphrasing existing material it was trained on (ex. "write about a new attack formation in soccer that allows through balls on offense").

Our dataset did have the following shortcomings that may have restricted our research:

1. **Simplicity:** Many of our prompts ended up being relatively straightforward for ChatGPT to answer. This limited our ability to test ChatGPT coherence when dealing with very difficult prompts. This is especially noticeable in the "Operating Systems" domain.

2. **Consistency:** While most of our prompts were promoting explanatory or synthesis outputs, "Fermi Questions" asked ChatGPT for mathematical reasoning that our coherence models were not equipped for. This resulted in lower coherence results that aren't entirely fair to compare to other domains considering how different the prompts were. Across the other domains, prompts also varied in their wording and goals for the response, which may have contributed to research error as well.

6.2 Replicability

Replicating our experiment would be quite easy, provided that the same prompts were used. In fact,

our own research found ChatGPT coherence outputs to be quite stable, which was seen in our average coherence scores not changing much when the model was re-ran. That being said, it may be difficult to reproduce our results from solely the domains if the prompt generation techniques used do not match ours. In particular, the "Fermi Questions" domain dealt with very specific algebraic questions that would likely produce different coherence results if phrased differently.

6.3 Model limitations

One of the biggest research takeaways from this project is the limitations of coherence models and the lack of universal metrics for it. This is most evident in our results relating to Fermi Questions, which often scored a lot lower on the HuggingFace coherence model than the other topics. While it is possible that the Fermi Questions outputs themselves were less coherent, it is much more likely that the HuggingFace model scored them a lot lower than others because of the heavy use of mathematical symbols in the responses. This is something that the HuggingFace model was not trained to follow, since it was trained on a database of Wall Street Journal articles.

The issue with the coherence model on Fermi Questions topics raises the question of how one might design a model for those sorts of problems. In theory, the process used to train the HuggingFace coherence model could be re-used to train a model on algebraic output. However, because the HuggingFace model is relative (only meaningful when compared to other results from the same model), it would still be difficult to compare these results to those from other domains if they were trained on a different model. All of this goes to show that there is still research to be done on finding useful objective metrics for coherence, and on modelling coherence in a way that is applicable to different types of text.

6.4 Ethics

As our research was done entirely on ChatGPT generated output, there is no potential for harm on individuals or specific social groups. However, because our research involves ChatGPT, a widely used LLM, it indirectly deals with some of the ethical issues that arise related to ChatGPT or the NLP field as a whole. For instance, coherence and readability metrics are very useful when discussing the usefulness of ChatGPT in education, particularly in determining reading level (readability), or in highlighting the perceived validity of generated output (coherence). Thus, incoherent or unreadable output could effect the usefulness of ChatGPT in these areas, and coherent but incorrect output (seen occasionally in our results) can result in content that falsely appears trustworthy.

7 Conclusion

Our project was to determine the overall coherence and readability of ChatGPT output when prompted across varying topics - from technical questions to creative strategies. We found that generally, ChatGPT remained coherent across domains. The outputs that included more logical reasoning or mathematical notation generally scored lower in coherence, likely due to the coherence model's limited training set. Future work could look universal coherence modeling across domains, or do a deeper study into prompts that result in lower coherence.

References

Aisingapore/coherence-momentum · hugging face.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Comput. Linguist.*, 34(1):1–34.

Carmine DiMascio. 2021. py-readability-metrics. GitHub repository.

1911-1986 Flesch, Rudolf. 1943. *Marks of readable style : a study in adult education*. Teachers College, Columbia University, 1943, New York (State).

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221–233.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

William A. (William Anderson) McCall. 1925. *Standard test lessons in reading*. Columbia University, New York City.

Selva Prabhakaran. 2022. Topic modeling in python with gensim.

Dat Tien Nguyen and Shafiq Joty. 2017. A neural local coherence model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1320–1330, Vancouver, Canada. Association for Computational Linguistics.