Measuring the impacts of parts of speech:
A method for evaluating input impact in a language model

Colin Kavanaugh
School of Computer science
North Carolina State University
cwkavana@ncsu.edu

Abstract

The continued proliferation of language models across new use cases has displayed one of the most common struggles of many neural network models. As language models become more adept and applicable in new contexts the issue continues to arise that they are uniquely inept at generating human understandable reasoning for their responses. This can make it very difficult for users and developers to understand how these models may adapt to their use case, specifically in contexts that do not have a direct level of truth that can be assigned to outputs. This paper seeks to identify a method for evaluating the reasoning of models in non-objective use cases through evaluating the "impact" of parts of a prompt. This is done through a test example whereby the importance of specific parts of speech will be measured and valued based on the level of change that occurs in a response when they are altered. Through this multiple differing methods will be analyzed and compared as potential metrics, and evaluated for this specific use case. Finally a conclusion will be drawn as to the importance of what metrics should consider, and how these methods can be applied in future work.

1 Introduction

Language models are unique due to their ability to represent the highly complex and context based meanings behind language in the way that humans tend to understand it. They are able to transform large amounts of contextual data into relationships that can appear as "intuitive" in a way that appears to mirror how humans approach language. This is what defines their usefulness, but it is because of their general applicability that makes them appear to work as a "black box".

There are multiple different methods used by developers to analyze models despite their complexity, but analyzing what specific associations exist does not necessarily improve the explanability of the model, especially in cases where a ground truth value does not exist. This paper seeks to use prompt modification as a method of identifying how models change when certain information is modified as a method of determining the importance a section of a prompt has on a model's output. By understanding the specific parts of a model that lead to certain results it may be possible to understand the relationships and associations that a model uses to formulate its outputs. If done on a scale that can represent the total space of prompts that a model is trained for it may be possible to map an understanding of how the model understands the prompts it is given, and the weight that it gives to them. This paper seeks to demonstrate this approach on an example case involving measuring the importance of parts of speech.

2 Goals and approach

Our goal is to evaluate the "impact" that four specific parts of speech have on a particular language model, and assign a value that represents the weight each piece is primarily given. Multiple metrics will be proposed for generating these values, and their results will be compared to find which value is most valuable for the specific case of measuring the weights of parts of speech. With this we hope to demonstrate the potential applications of using prompts as a method for evaluating model “impact”, and will argue the direction of future work and cases where this approach is most appropriate.

3 Experiment setup

The goal of our experiment will be to systematically alter a set of base prompts for each part of speech and measure the resulting change in the model's response. By measuring the change in output from the origin prompt and the modified prompt we can view the “impact” that the change caused in the model's output, and this conclude on the importance the modified word had on the model's understanding of the prompt. By modifying each each part of speech in set ways across a set of base sentences we can then find the total percentage of changes made by each part, and thus reach a conclusion as to the importance that the evaluated model grants to that part of speech. For discussion on the limitations of this experiment, see section 5, and for discussion on future potential expansions of this experiment, see section 6.

3 .1 Prompt generation

Each step of the experiment involves a set of base sentences and for this experiment 9 unique sentences were generated, each based off of prompts from one of 3 model applications involving no expected correct values, those being review generation, conversation response, and customer support.

In each sentence one instance of the the following four parts of speech were identified, noun, verb, adjective, adverb. For each part of speech a set of syntactically associated words were generated with the use of the Wordnet lexical database. Wordnet associates words together in a tree based structure of associated “synsets” of shared concept or meaning. These syntactic similarities varied per part of speech and were as follows

Noun:
Hypernym-A parent in the base word's synset. A more generic term.
Hyponym -A child in the base word's synset. A more specific term.
Sister term-A word who shares a hypernym relationship with the base word.
Verb:
Troponym-A word who's meaning is the highest of its associated synset tree.
Hypernym-A parent in the base word's synset. A more generic term.
Adjective:
Synonym-A word with equivalent meaning
Antonym-A word with opposite meaning
Adverb:
Synonym-A word with equivalent meaning
Antonym-A word with opposite meaning

Each of these parts of speech were also switched for a random word of the associated type, which was referred to as the the “unrelated” term. With these changes each sentence was modified into 13 unique variations that were each applied to the model.

3 .2 Model representation

The Model used in this experiment was the pretrained Llama 2 7 billion parameters model from hugging face. The model would be given one sentence at a time, and the hidden state of the model would be saved. The hidden states were output as a tuple of 33 tensors with a size of [1,N,4096], where N was variable based on the input prompt. It was run on the NCSU Root Cluster(ARC) on the AMD Epyc Rome nodes.

3 .3 Comparison metrics

Once each output was generated, the difference between the two outputs was then measured with 3 distinct methods, each of which gave a single numerical value to represent the total change from the first distribution to the second.

1. Average point L2 norm: The total set of points from an output is averaged to find a single point that represents the position of the average point of the set. The average point of the base sentence and the compared modified sentence were then measured with an L2 norm which represents the distance between two vectors and can be represtned by the following equation.

(1)(||X||) = √(x1² + x2² + … + xn²)

2. Point to Point Norm: An L2 norm calculation was made for each point in each output to the other one associated point in the second output. For example the base output(O) node[0] was compared to the altered output(O`) node [0], for each node. Due to the model producing 33 output point, this would produce 33 norm results, which would then be summed to calculate the final value.
3. Jensen Shannon Divergence: Involves preforming two Kullback-Leibler Divergence calculations and taking the average of the two results. The Kullback-Leibler divergence is a calculation that compares two statistical representations by calculating the amount of entropy needed for one representation to be equivalent to another. It can be represented with by the following equations
(1)KL(P||Q) = Sum(P(x)log(P(x)/Q(x)

(2)JSD(P||Q) = 1/2 * KL(P||M(+ 1/2 * KL(Q||M)

Each of these methods were implemented with the assumption that no one method was perfect and each would represent diffent aspects of the data. The idea is that each may be potentially useful for specific use case, and the goal was to identify which one would be best utilized for the case being examined here. The Average Point norm calculation was used as a baseline calculation, as it is a fairly simple calculation and is used in other areas of machine learning to calculate the difference between clusters. It was however expected to lose information relating to the overall distribution and shape of the output due to compressing the data into a single point.

The Point to Point comparison was chosen as a potential method to measure the shape of a distrubtion as it can represent the total difference. It does however work under the assumption that each point has a defined associated point in the other ouput, which may be the case in nearly identical distributions, but is expected to result in near random distrubtions as output become less associated.

Finally the Jensen Shannon Divergence was chosen as another method to analyze the overall shape of the model and to factor in more aspects of the distribution than the Average Point Norm. The method involves representing the output of the model in as a statistical representation, and as such is not a direct distance measurment, but instead a mathematical description of difference. This is useful for this context as it can describe aspects that the simple distance measurments lose. Being a more complex calculation however it is more computationally intensive than the previous two calculations, especially within the context of the high dimentional values that the outputs for the model were in. An additional statistical comparison method called the Wasserstein Divergence was an additional method that was attempted, but proved to be too intensive for the data that was being worked with due to the high dimentional space of the output data. When compared to the Kullback-Leibler Divergence however it is additionally less prone to outlier values, and so was chosen for this use case.

Each of these 3 metrics produced a single value for each comparison that were then summed together and normalized to determine a final “impact” value for each part of speech, which could then searve as a final value.

4 Observation & analysis

The results for each test were sorted by syntactic change and measured by the average and median values between all the sentences tested. Additionally the average values for each part of speech were calculated across syntactical changes to determine the final percentage weight that each part of speech had for each method. The details and discussion of observed results are bellow.

4.1 Average Point

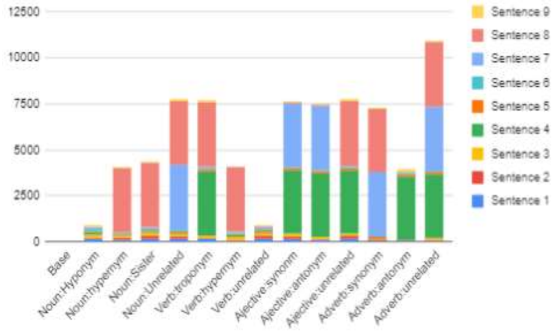

Figure 1: **Summed values for Average Point distance by syntactic change from all sentences.**

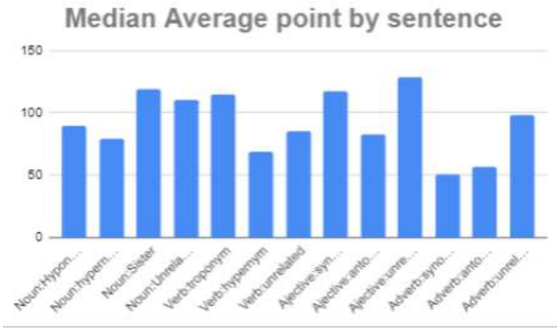

Figure 2: **Median values for Average Point distance across all sentences by syntactic change**

The Average Distribution is heavily influenced by outliers

The most striking feature of the summed values table is that the total values are defined by a set of outliers which made up a majority of the average scoring. This was a surprising observation as the Average Distance was expected to be the most stable method, and its issues were predicted to be in measuring certain values as too similar rather than as too distinct. We can additionally see that the ordering of the "impact" of syntactic changes was significantly different between the median and summed values, and this indicates that the results are not reliable or stable. For this reason the median values are seen as the most representative measurement for the methods distribution. This could be due to multiple factors, including the small amount of data being used, the model, or the method itself, but not enough testing was done to diffidently conclude the cause. More of this topic will be discussed in sections 5.

The median Values are close in scale across sentence changes

Compared to the summed values and especially compared to the median charts of the other methods, the values shown on the Average Point medians are all relatively very similar, with the lowest and highest values having a difference of around a 250% scale. This is notable as it was expected that this method would lead to relativity similar results relative to other methods and this appeared to be the case.

4.2 Point to Point

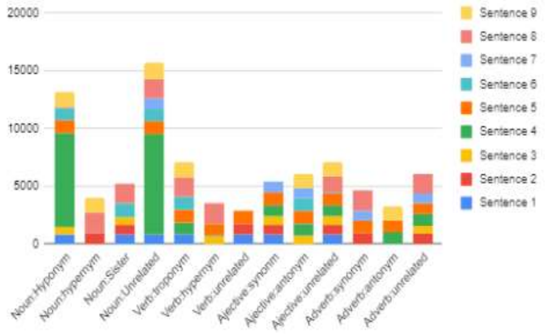

Figure 3: **Summed values for Point to Point distance by syntactic change from all sentences.**

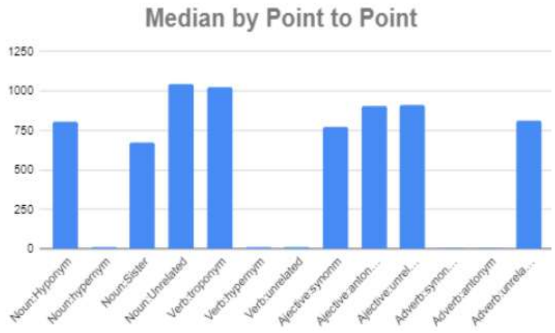

Figure 4: **Median values for Point to Point distance across all sentences by syntactic change**

The Median chart is similar in rankings, but not in scale

The most notable feature when analyzing the point to point graphs is the large change in scale between the lowest "impact" changes and the highest between graphs. While the summed graph does appear to have at least 2 instances of large outliers, the overall ordering of syntactical changes, especially within parts of speech, were consistent between the median and average values. This indicates at least a level of reliability in the data when it comes to viewing its consistency. What is unique is in the median value charts there is a set of 5 syntactical changes that have values that are close to 0, with the scale between the lowest and highest median values being upwards of 20500% difference. These were all the values that were shown to be the least “impact” both on the summed chart, and from other models, but the difference in scale with the Point to Point method is very unique.

A comparison can be drawn to the distributions of models that could be considered “overfit” in other contexts, and while it is not an applicable term in this case, it was expected that the point to point method would result in data with the widest gaps between measurements. It cannot be concluded if what specifically is causing this type of distribution, and future work would be necessary to conclude.

4.2 Jensen Shannon Divergence

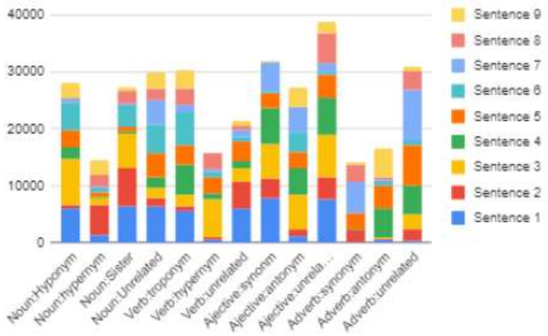

Figure 5: **Summed values for Jensen Shannon Divergence by syntactic change from all sentences.**

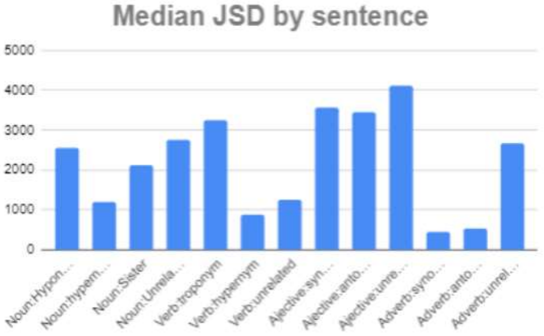

Figure 6: **Median values for the Jensen Shannon Divergence across all sentences by syntactic change**

Median and Summed values aligned
The JS divergence appeared to show the most stable distributions between both the median, with both the ordering of terms being similar, especially between parts of speech, as well as the ratio between the smallest and largest values being within a relative scale of one another (250% for summed vs 800% for median). It is important to note that the scale of low “impact” changes were more pronounced by the median value, indicating that there are a set of larger values for each lower “impact” changes that skewed the sum and average scores towards closer values.

Showed best value distribution for measuring parts of speech
As previously mentioned the goal of the project is to identify the average weight that each part of speech causes, and so the ideal method would bias towards methods that favor regular cases over factors such as outliers. Additionally the method should be able to detect small changes as the expected measurement is likely to be potentially very similar. With those criteria in mind it appears that the Jensen Shannon Divergence is the most reliable method for measuring these values in this case, as it shows a consistent distribution between the average and median that indicates a resistance to outlier bias, as well as measurements that can appear to identify small changes in very similar prompts.

4.5 Syntactic change impacts

Syntactical impacts were similar across methods
Across all methods, with the only exception of summed values with Average Point distancing, the ordering of syntactical “impact” was very similar, especially within parts of speech. It was observed that the for each part of speech the hypernym was found the be the lowest “impact” change across methods for the nouns and verbs, and synonyms for the adverb. Similarly the unrelated term was found to be the highest “impact” change for the nouns, adjectives, and adverbs across methods. Points of interest appeared to be with the lowest “impact” change for adjectives, and with the

highest “impact” change for verbs. In all cases the highest “impact” term for verbs was the troponym. This was surprising as the troponym was expected to be related to the original word, and therefore give more similar values, especially compared to the random unrelated terms. The reasoning for this result is not clear, as it could be a result of the model being tested, the small sample size of the data, or could be related to the nature of Wordnet's synset organization.

The lowest impact adjective changed depending on the measuring method

The largest disagreement that the methods had was on the least impactful adjective, as the Average Point distancing had different ordering for median and summed, the Point to Point distance found the synonym to be the least impacful, and the Jensen Shannon Divergence found the antonym to be the least impactful change. It was expected that the synonym would be the most similar to the base word, and therefore would lead to a smaller change, but due to the Jensen Shannon Divergence being concluded more reliable for this case, its results would likely be trusted over the Point to Point Distance. Regardless, this being the only ordering of syntactic changes that was different between methods makes it notable, however definite conclusions as to the cause were not found.

4.6 Part of Speech Impact

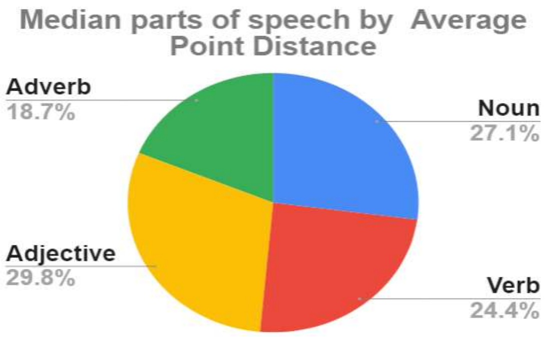

Figure 7: Average impact of each part of speech by average syntactical change **using Average Point Distancing**

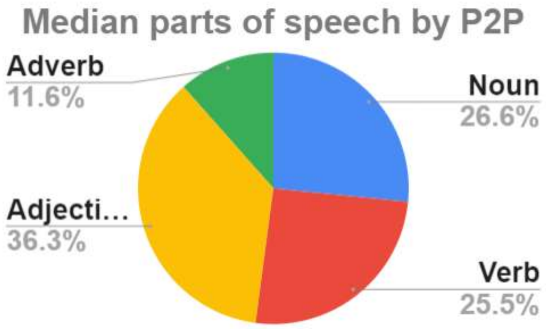

Figure 8: **Average impact of each part of speech by average syntactical change using Point to Point Distancing**

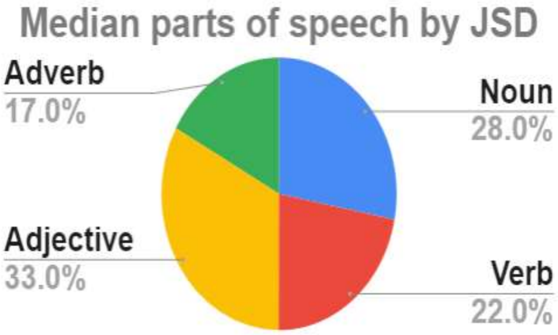

Figure 9: **Average impact of each part of speech by average syntactical change using Jensen Shannon Divergence**

Impact order for parts of speech was consistent with median value measurements, but not with summed.

It was found that the ordering of parts of speech by “impact” was consistent between the median values, finding the adjectives to be the most impactful, followed by nouns, verbs, and with adverbs being the least impactful part of speech. This was found by the median measurements for all methods, as well as the summed values of the Jensen Shannon Divergence. The summed values for the Average Point distance and Point to Point distance found both different highest “impact” as well as different lowest “impact” parts pf speech. For these reasons it was concluded that this was likely the most accurate representation of part of speech “impact”.

5 Limitations and future work

The results observed in this paper are heavily limited by the lack of comprehensive investigation of all the factors that lead to the results, and due to this all observations are not given a definite conclusion. Future work will be needed to confirm and determine the validity for any results seen.

5.1 Model limitations

The Llama-2 7b model used in this paper was chosen due to being able to be local run with limited computational resources, and does not reflect the wider scale of language models. All test conducted are valid only for the specific model being tested. Future work will require testing with a greater number, and higher complexity models to draw conclusions for a wider use case.

5.2 Data Limitations

The sentences picked in this paper were chosen as a simple representation of a potential case with no definite correct response, it does not reflect any specific application or use case, and a proper dataset would be necessary to draw any conclusions about the results. Future work would have to identify the exact use case being meet, and develop a set of data that can accurately represent that use case for its results to be conclusive.

5.3 Syntactic limitations

The syntactical changes used in this case were chosen due to their universal application with the Wordnet lexical database, and were limited to the connections mapped there. They are not a representative example of every syntactical similarity that can be found, and future experiments would likely make use of a wider selection of sentence alterations that can accurately represent a problem space.

5.4 Validation limitations

The values found for each method of comparison were unable to be determined as accurate due to a lack of verification method. Results were limited to being compared and analyzed, and

The applicability of the work is questionable due to the lack of comprehensiveness. Future work may benefit from a measure of human verification, specifically when it comes to outlier or extreme values in order to measure some form of accuracy for method measurements.

5.5 Test Limitations

The observations found in this paper relate to a small test case in order to try and show the usefulness of "impact" measurements in NLP, and do not currently represent a conclusive investigation of any particular use case. The choices made in terms of measurement methods and priorities were made for a test case that does not represent any specific application. Future work would benefit from defining a specific application for the use of "impact" measurements in order to properly demonstrate how it can be used to better understand NLP.

6 Conclusions

The use of prompt "impact" as a measure of relevance parts of speech can be useful as a measure of the inner dynamics of a model, but requires a more definitive application and future testing. An example experiment was preformed however its total, but it was concluded that the methods used were not conclusive and require future work. Future work will require using a larger pool of models, as well as testing on higher complexity models. The test data will also need to be expanded to properly represent the types of prompts a model would be used with in the scope of the use case, and so smaller use cases would work better as an example case. With these factors accounted for the use of an "impact" measure would be a useful tool for improving the understanding and interoperability of language models using NLP.

7 References

Fellbaum, C. (2010). WordNet. In: Poli, R., Healy, M., Kameas, A. (eds) Theory and Applications of Ontology: Computer Applications.Springer,Dordrecht. https://doi.org/10.1007/978-90-481-8847-5_10

Melville, P., Yang, S.M., Saar-Tsechansky, M., Mooney, R. (2005). Active Learning for Probability Estimation Using Jensen-Shannon Divergence. In: Gama, J., Camacho, R., Brazdil, P.B., Jorge, A.M., Torgo, L. (eds) Machine

Learning: ECML 2005. ECML 2005. Lecture Notes in Computer Science(), vol 3720. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11564096_28

Hackmann, Stefan, et al. “Word Importance Explains How Prompts Affect Language Model Outputs.” *ArXiv (Cornell University)*, 5 Mar. 2024, https://doi.org/10.48550/arxiv.2403.03028. Accessed 23 Apr. 2024.

-