DS-GA 1016: Computational Cognitive Modeling Comparison of Neural Network and Human Next Word Predictions in Article Titles

Yasi Asgari (ya2193@nyu.edu)
Nikki Gharachorloo (nng235@nyu.edu)
Alex Herron (ah5865@nyu.edu)
Victoria Xie (xx2179@nyu.edu)

Introduction

In recent years, next word prediction has become an increasingly popular modeling task. Not only is next word prediction useful for tasks like assisted typing, but it provides larger insight into how language models understand textual data. Language models compute a probability distribution of all the potential choices for the next word, then output the word with the highest probability for the given context. These choices of most likely words and their corresponding probability distributions provide some intuition into how these language models work, which is particularly helpful given the black-box nature of neural networks. Additionally, a study from neuroscientists at MIT proposes that the function of these language models can also provide an improved understanding of how the language-processing centers of the human brain work.

Our project focuses on the comparison of two different neural network models and human participants for the task of next word prediction within the context of article titles. The natural language processing (NLP) techniques used for next word prediction are similar to how the human brain would predict the next word in a given context for a number of reasons. Both NLP models and the human brain use contextual understanding to predict the next word, meaning they take into account the words that come before the target word, as well as the overall context of the sentence or passage. Another similarity is that they both learn from experience. NLP neural nets are trained on large amounts of text data to learn patterns and make predictions based on those patterns, while the human brain learns from experience and exposure to language over time, then uses that knowledge to make predictions about what words are likely to come next. NLP models can be trained on a set of labeled data to adjust their predictions based on the correct answers, while the brain can adjust its predictions based on feedback from the listener or reader.

For this project, we gathered useful background information for language modeling from ”Deep Learning for Natural Language Processing: A Comprehensive Review” by Li Deng and Yang Liu. This book provides a exhaustive overview of deep learning methods for natural language processing, including topics such as language modeling, machine translation, and sentiment analysis. The authors provide an overview of the different architectures and techniques used in these methods, as well as their strengths and limitations. This book covered various deep learning methods used in natural language processing, and encouraged us to use an LSTM for our first model. This book also highlights the importance of understanding the strengths and limitations of these methods, which can help guide the development of more effective models.

Additionally, we drew inspiration from the paper ”Human-level concept learning through probabilistic program induction” by Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. This study explores a computational approach to concept learning that is inspired by human cognition. The authors present a computational model called Bayesian Program Learning (BPL) that aims to capture human-level concept learning abilities. BPL represents concepts as basic programs, and uses a Bayesian criterion to explain observed examples. It was evaluated in a one-shot classification task using handwritten characters from different alphabets. Interestingly, the model reached human-level performance, using classification accuracy as the evaluation metric (calculated as the log posterior predictive probability). This encouraged us to compare our neural network models to human participants. Additionally, contextual understanding played an important role in various aspects of this paper, including image generation and visual relationship parsing. By incorporating contextual cues, their models were able to generate visually meaningful concepts that aligned with descriptions. This inspired us to closely study the context for various next word predictions tasks, namely for the subset of 10 questions answered by human participants. Finally, this paper emphasized the fact that machine learning methods are currently limited in their capacity to learn new concepts from sparse data, and the ability to learn flexible representations.

Lastly, we gained a broader background on the task of next word prediction from ”Next word prediction using Deep Learning: A Comparative Study” by M. Soam and S. Thakur. This paper covered a broad array of different deep learning architectures used for next word prediction, including N-gram modeling, convolutional neural networks, and recurrent neural networks. Additionally, this paper briefly summarized many other papers dedicated to next word prediction, some of which were focused on aiding user typing speed. In addition to the widespread use of LSTMs for next word prediction, this paper highlighted the extensive success of such models in various domains.

Methodology and Modeling

Publication	Number of Articles
The Startup	2569
Towards Data Science	1245
Data Driven Investor	676
UX Collective	473
The Writing Cooperative	330
Better Marketing	213
Better Humans	26

Table 1: Number of Articles for Each Publication in Dataset

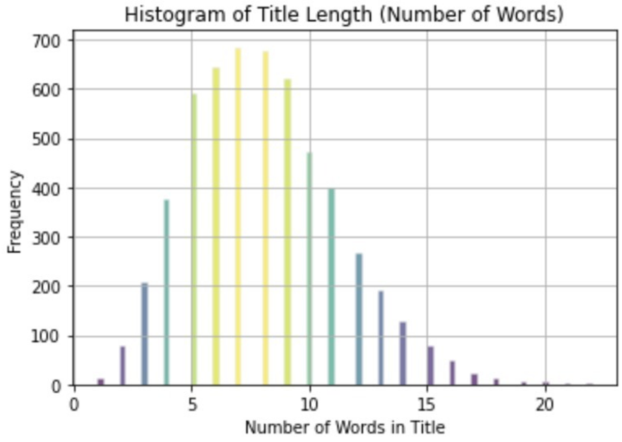


Figure 1: Histogram for Title Length

Our approach can be broken down into 3 steps. We first trained a long short-term memory (LSTM) model, then made next word predictions on test data using both the LSTM and a pre-trained Transformer model, and finally compared the models' ability to predict next words with results from human participants on a smaller subset of the test data. The LSTM was trained on a dataset that includes information from 6508 medium-length published articles collected in 2019, from 7 selected publications. The distribution of these various publications can be seen in Table 1, highlighting the fact that many of these titles are related to startups and data science. The LSTM was trained using an 85:15 train/test split.

LSTMs are a type of recurrent neural network (RNN) that is often used in natural language processing tasks such as next word prediction. In an LSTM model, the input sequence is one-hot encoded, passed through an embedding layer which maps each word vector to a dense embedding vector of fixed size, and then fed into an LSTM layer. The output from the LSTM layers is passed through a fully connected output layer with an activation function to generate the probability distribution over the vocabulary for the next word. During training, the model is fed text from article titles, and then asked to predict the word following that sequence with the highest probability. For our LSTM, we used softmax as our activation function, categorical cross entropy as our loss, and trained for 50 epochs (requiring under an hour locally). Throughout training, accuracy, loss, and recall were calculated, which can be seen in Figures 2, 3, and 4, respectively.

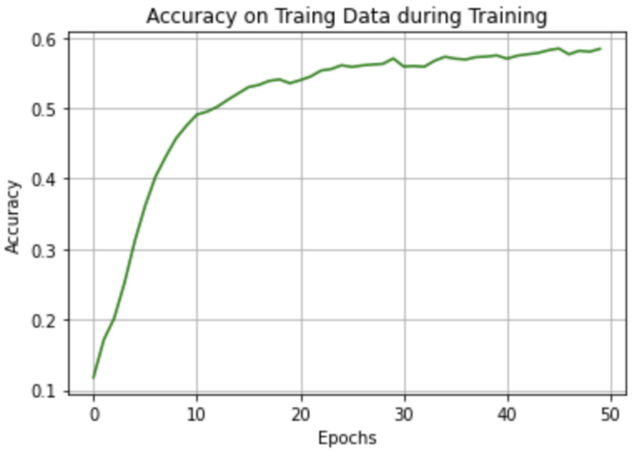


Figure 2: Training Accuracy Curve for LSTM

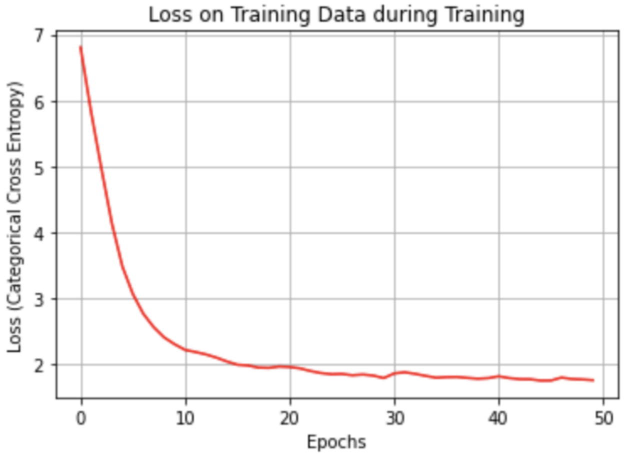


Figure 3: Training Loss Curve for LSTM

The Transformer model is another type of neural network architecture that uses self-attention to model the relationships between words in a sequence. The input sequences are first encoded and embedded, then passed through several Transformer layers, which use self-attention to compute a weighted sum of the token embeddings, where the weights are computed based on the similarity between each pair of tokens in the sequence. This allows the model to capture complex dependencies between the words in the sequence. The output of the final Transformer layer is then passed through a linear layer with softmax activation to generate the probability distribution over the vocabulary for the next word. For our project we used the pre-trained Transformer model T5 (Text-

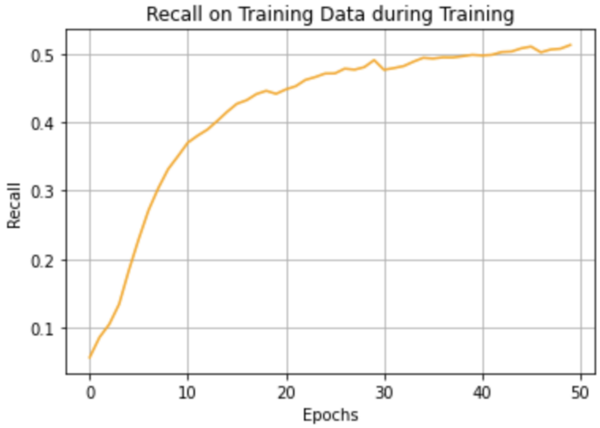


Figure 4: Training Recall Curve for LSTM

to-Text Transfer Transformer) developed by Google AI Language, and made available by Hugging Face and the Allen Institute.

LSTM and Transformer models are both good for next-word prediction due to their ability to capture long-term dependencies in sequences and model the relationships between words in a sequence. The key feature of LSTM is its memory cell, which is able to store information about the past input data and selectively retain or forget this information based on the current input. This allows the LSTM model to capture long-term dependencies in sequences, which is important for next-word prediction since the prediction of the next word often depends on the words that came before it. On the other hand, Transformer models are based on the self-attention mechanism, which allows them to capture the relationships between all words in a sequence in a parallel manner. This allows the Transformer model to capture complex dependencies between words, including long-term dependencies, without being constrained by the sequential nature of RNNs. LSTM models are often used for smaller datasets or when more precise control over the sequence is required. In contrast, Transformer models are used for large scale problems, when the sequence is long and the self-attention can be used to capture important dependencies.

In the last stage of the study, we randomly sampled 10 examples from the test set and asked 15 human participants to predict the last word for each title. For each item, participants were given 10 options. In the case that the LSTM predicted the next word correctly, these 10 options were the 10 most probable next words for each example produced by the LSTM model. If the LSTM was incorrect, the options included the top 9 options from the LSTM in addition to the correct word. For example, in question 1, "How Small Beat Big ____", participants were given all the words in the title except the last word and instructed to choose from the list of ten possibilities to complete the title, which are: "writer's", "form", "beats", "big", "outcomes", "pipelines", "predicting", "benefit", "myths", and "obsessed", respectively. 15 valid responses, each containing answers to 10 items, were collected from these participants.

The objective was to compare neural network predictions to human responses, as we were curious to see the imitations of these models in terms of their understanding of the meaning and context of the words they are predicting, and ultimately compare these predictions to those of human participants. We want to derive insights on questions such as: does the underlying function of language models resemble the function of language-processing centers in the human brain? Do the language models exhibit signs of understanding of the human language? What are some limitations of language models?

Results

First, we compared the results for the LSTM trained on article title data and the pre-trained Transformer. Results for these two models evaluated on the test data can be seen in Table 2. The Transformer model performed better than the LSTM for accuracy (over 2% higher) and precision (over 0.002 higher), and tied with the LSTM for recall. Additionally, it should be noted that the highest accuracy was still below 10%, highlighting the inherent difficulty of this task.

Model	Accuracy	Precision	Recall
LSTM	0.0747951	0.02315	0.02664
Transformer	0.0973361	0.02585	0.02664

Table 2: Results for Models and Human Responses on All of Test Data

Next, we compared the results of the LSTM, Transformer, and aggregate human participants on the subset of 10 questions from the test data. The results of these different models/participants can be seen in Table 3. This table highlights that both the LSTM and Transformer only had an accuracy of 10% (meaning they only answered 1/10 questions correctly), while the aggregate human responses had an accuracy of 59.3% (with 89/150 responses being correct).

Additionally, the various questions used in this subset of the test data, target responses, LSTM predictions, Transformer predictions, and the majority of survey responses can be seen in Figure 5. Question 6 was the only question correctly predicted by either model, while there were many questions where the majority of participants got the correct answer, such as questions 1, 3, 4, 5, 6, 8, 9, and 10.

Clearly, even with such a limited amount of human data, humans appear to far outperform these models in next-word prediction. It is important to note, however, that the tasks performed by these models and humans are fundamentally different. While humans were given a choice of 10 potential answers, these models were not. Rather than choosing from 10 options, these models had to choose from every single one of the words in their vocabulary, which is an inherently more difficult task. It is likely that human participants would've had a significantly lower aggregate accuracy if they were not given multiple-choice options.

Model	Accuracy
LSTM	0.1
Transformer	0.1
Human Responses	0.593

Table 3: Results for Models and Human Responses on Subset of Test Data

In order to better understand how the words selected by survey participants compare to the predictions made by the LSTM and Transformer models, questions 4, 5, 6, 7, and 9 can be further studied. Questions 4, 5, and 9 highlight cases where many participants chose the correct word, whereas most participants chose incorrectly for question 7. Lastly, question 6 was the only question where either the LSTM or Transformer chose correctly.

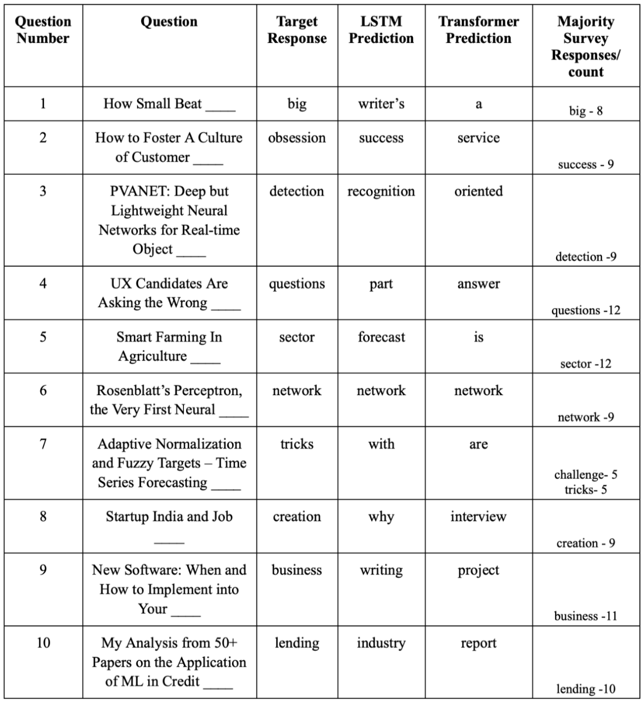
For question 4, one of the potential reasons “questions” was a popular choice among the survey participants (chosen by 80% of participants) is due to contextual knowledge. It makes intuitive sense that candidates would ask the wrong ”questions” in this context. Additionally, the bi-gram ”asking questions” is relatively common, which would be a strong indicator for both human participants and models. However, despite ”asking questions” being a common bi-gram, this did not occur very frequently in the article title training corpus. The LSTM and Transformer chose ”part” and ”answer,” respectively, which does not make sense in the context.

With question 5, it can be argued that ”sector” was primarily selected in the survey responses (also chosen by 80% of participants) because ”agriculture sector” is another common bi-gram. The LSTM chose ”forecast,” the second most common response among participants, while the Transformer model chose ”is.” Although the Transformer was pre-trained extensively for next-word prediction, it was not trained to explicitly predict the last word in an article title. Consequently, it chose a word which does not make sense in the context of the last word of the given article title prompt.

In the case of question 9, it is clear that the options that are verbs do not fit grammatically with the prompt. This leaves ”writing” and ”business” as the only two choices that make sense. Business was chosen correctly by the majority of responses (73.3% of participants). This highlights the importance of contextual grammar for participants answering multiple-choice questions. The LSTM and Transformer chose ”writing” and ”project,” respectively, which were not correct despite being the correct part of speech (noun).

Next, question 7 appeared to be the most difficult question for humans, with only 1/3 of participants answering correctly. This is most likely because two of the options (”challenge” and ”tricks”) make sense in this context. However, although both of these two potential options make sense in the context of the prompt, only one of them is the correct choice. The LSTM and Transformer selections of “with” and “are,” respectively, are poor choices because they require an additional word despite being the last word in the article title.

Finally, the LSTM and Transformer only answered question 6 correctly by choosing ”network.” The LSTM was trained on the titles of many articles related to deep learning topics, many of which mentioned neural networks, which likely lead to this correct prediction. The transformer, while trained on a larger and broader training data, also identified this common bi-gram, likely using the context of ”perceptron” as well as the preceding word.



Question Number	Question	Target Response	LSTM Prediction	Transformer Prediction	Majority Survey Responses/ count
1	How Small Beat ____	big	writer’s	a	big - 8
2	How to Foster A Culture of Customer ____	obsession	success	service	success - 9
3	PVANET: Deep but Lightweight Neural Networks for Real-time Object ____	detection	recognition	oriented	detection -9
4	UX Candidates Are Asking the Wrong ____	questions	part	answer	questions -12
5	Smart Farming In Agriculture ____	sector	forecast	is	sector -12
6	Rosenblatt’s Perceptron, the Very First Neural ____	network	network	network	network -9
7	Adaptive Normalization and Fuzzy Targets – Time Series Forecasting ____	tricks	with	are	challenge- 5 tricks- 5
8	Startup India and Job ____	creation	why	interview	creation - 9
9	New Software: When and How to Implement into Your ____	business	writing	project	business -11
10	My Analysis from 50+ Papers on the Application of ML in Credit ____	lending	industry	report	lending -10

Figure 5: Responses for Subset of Test Data

Closely examining these 10 prompts highlights the fact that both the LSTM and Transformer consistently predict words that do not make sense in the context of the last word of an article title. This can be seen more closely in Tables 4 and 5. In Table 4, the top 3 words predicted by the LSTM were ”the”, ”and”, and ”a”; none of which make grammatical sense at the end of an article title. Similarly, in Table 5, the top 3 words predicted by the Transformer were ”a,” ”is,” and ”the,” which also do not make sense grammatically at the end of an article title. This is to be expected for the Transformer, which was pretrained on a very large dataset that was not specific to article titles. However, this is not necessarily expected for the LSTM, which was explicitly trained for next word prediction in article titles.

Additionally, Tables 4 and 5 highlight the fact that the Transformer was more likely to choose a single word (”a”) for many predictions (120), whereas the LSTM had a more flattened distribution of word predictions (i.e., the top word only had a frequency of 42). This points to the fact that the Transformer was pretrained to learn broader trends in many facets of English text, which ultimately results in less diverse predictions. It is likely that, if the LSTM was trained on a larger dataset, its distribution of word predictions would trend more closely to those of the Transformer.

Word predicted by LSTM	Frequency
the	42
and	24
a	20
in	18
you	14
your	14
strong	12
with	10
to	9
learning	9

Table 4: Top 10 Words Predicted by LSTM

Word predicted by Transformer	Frequency
a	120
is	94
the	34
of	16
learning	15
and	15
how	14
are	13
work	10
to	10

Table 5: Top 10 Words Predicted by Transformer

Humans are able to use their understating of context to very skillfully predict the next word in a sequence, despite the fundamental difficulty of the task. Understanding linguistic elements like grammar and syntax as well as cultural and situational elements like tone, genre, and audience may fall under the contextual understanding category. In addition, due to their flexibility, humans are able to produce predictions that may not be only based on statistical patterns in the data by utilizing their ingenuity and adaptability. Also, humans have the ability to re-evaluate their responses if their initial choice doesn't seem correct, while models are not trained to re-evaluate their results.

In conclusion, contextual knowledge, words that are frequently used together, and grammatical rules can be helpful for next-word prediction for humans, but since models do not comprehend these ideas on the same levels as humans, they do not perform with high accuracy compared to the accuracy of the survey participants.

Discussion/Conclusion

For this project, we aimed to investigate how language models perform compared to humans on next-word prediction tasks, specifically with article titles. To study this, we trained an LSTM model on a set of article titles, and use a pre-trained Transformer model. Both of these models were evaluated on a test set of the article titles. We then evaluated 15 human participants on 10 questions randomly selected from the test set, and compared the human results to the language model results.

The results of our study suggest that language models have some similarities with human cognition, such as their use of contextual understanding and learning from experience. However, the models still have limitations in understanding human language and making accurate predictions. Our comparison of the LSTM and pre-trained Transformer models showed that the Transformer model performed better in predicting the next word in an article title. However, human participants outperformed both models by a substantial margin.

One possible explanation for the difference in performance between the models and humans is that the models lack a broader understanding of the world that humans possess, which can result in challenges with understanding sarcasm, irony, or other types of figurative language that rely on contextual understanding. Humans have a vast store of knowledge, experience, and common sense that helps them understand language in a way that is currently beyond the reach of language models. While these models are capable of learning specific patterns from the training data, they lack the capacity to draw on experience or common sense in the same way that humans do. For example, in item "How Small Beat ____", more than half of the respondents chose the correct answer – "big", while the LSTM and the Transformer models predicted "writer's" and "a", respectively. It is safe to assume that human participants picked "big" as a contrasting adjective to the previous word "small" based on common sense, to convey an intriguing and attention-grabbing message to readers. The models, however, failed to make the correct prediction due to a lack of contextual understanding, and produced predictions solely based on long-term dependencies in the sequences.

Additionally, the models may be limited by the amount and quality of the data they are trained on, as well as the architecture and parameters of the model itself. For example, these models may have difficulty with uncommon or rare words that are not present in the training data. This issue of training data is more likely to limit the LSTM than the Transformer, as it was trained on a substantially smaller dataset. In contrast, humans can use their knowledge and experience to infer the meaning of such words even if they have never encountered them before. Ambiguity is another case in which it can be challenging for models to correctly recognize the meaning of a word that has multiple meanings depending on the context in which it is used, whereas it is a lot easier for humans to infer the correct meaning from the context.

It is also important to acknowledge the selection bias that is potentially present among the group of participants who completed the survey items. Some of the respondents are graduate students majoring in data science and related fields, who are familiar with and have a better understanding of published scientific papers. For instance, in item “PVANET: Deep but Lightweight Neural Networks for Real-time Object ____”, the majority of the participants chose “detection”, which is the correct answer, whereas the LSTM and the Transformer model predicted “recognition” and “oriented”, respectively. Even though both of the model predictions make sense semantically, a data science graduate student might be aware that “object detection” is a computer vision technique for locating instances of objects in images or videos. As such, the sample might not accurately represent the population, leading to a higher accuracy score.

Regarding future work, there are certainly many pathways to expand upon this topic. One clear next step would be expanding the training data used by the LSTM, and seeing how its performance would change. In addition to the accuracy of an LSTM trained on a much larger dataset, it would be interesting to see if the distribution of the LSTM’s predictions would become more similar to those of the Transformer as the size of the training data was increased.

Secondly, it would be useful to expand the scope of this study beyond article titles. Although we concluded that there are several reasons why humans perform better in predicting words in article titles (such as the use of grammatical context for final words), it would be very useful to compare the performance of humans and neural networks in next-word prediction in other domains.

Third, additional metrics could be created to provide deeper insight into the results of these predictions. Although the primary metric we used was accuracy, this doesn’t account for predictions that make sense in the context of the article title, but happen to be wrong regardless. Ideally, these predictions should be differentiated from those that don’t make sense semantically. One specific way to implement such an evaluation metric would be P.O.S. accuracy. This metric would measure the percent of predicted words matching the target word’s part of speech (e.g., nouns, adjectives, verbs). The use of such a metric would provide increased insight into what these models and human participants do well, and what they do poorly.

Lastly, in the interest of analyzing the results of models trained on large amounts of data, it would be helpful to see how Large Language Models (LLMs) perform on similar tasks. As a preliminary step in this direction, we asked ChatGPT to predict the next word for the same 10 prompts we asked our human participants. First, we asked ChatGPT with zero context to ”Predict the next word for each of these,” followed by each of the questions. Second, we asked Chat GPT to do the same task, but with the added context that ”these are all article titles related to start-ups, data science, and machine learning.” Interestingly enough, ChatGPT got 4/10 of the questions right when given zero context, and got 2/10 of the questions right when provided context about the types of prompts and their content. It is difficult to conclude why this added context decreased ChatGPT’s performance. However, it’s certainly worth noting that both the uninformed and informed versions of ChatGPT performed better than the LSTM and Transformer on this small subset of 10 questions.

In conclusion, our project provides insights into the similarities and differences between language models and human cognition in predicting next words. While both the LSTM and Transformer show promise in their ability to learn patterns from language data and make predictions, they still have substantial limitations that need to be addressed. Further research in this area could lead to language models better suited to word-prediction tasks, which may also provide insight into the cognitive abilities of humans.

References

- Lazar, D. (2020, June 30). Medium articles dataset. Kaggle. https://www.kaggle.com/datasets/dorianlazar/medium-articles-dataset?resource=download

- Allenai/T5-small-next-word-generator-qoogle · hugging face. allenai/t5-small-next-word-generator-qoogle · Hugging Face. (n.d.). https://huggingface.co/allenai/t5-small-next-word-generator-qoogle

- JB;, L. B. R. (n.d.). Human-level concept learning through probabilistic program induction. Science (New York, N.Y.). https://pubmed.ncbi.nlm.nih.gov/26659050/

- Soam, M., amp; Thakur, S. (2022). Next word prediction using Deep Learning: A Comparative Study. 2022 12th International Conference on Cloud Computing, Data Science amp; Engineering (Confluence). https://doi.org/10.1109/confluence52989.2022.9734151

- Mishra, A. K., Pal, S., Pal, S. K. (2016). Dynamic Lexicon Generation for NLP Applications. In Proceedings of the International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES).

- Gatys, L. A., Ecker, A. S., Bethge, M. (2015). A neural algorithm of artistic style. Journal of Vision, 15(12), 950-950.

- Schrimpf M, Blank I, Tuckute G, et al. The Neural Architecture of Language: Integrative Reverse-Engineering Converges on a Model for Predictive Processing. BioRxiv. 2020:2020.06.26.174482. doi:10.1101/2020.06.26.174482

- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., amp; Liu, P. J. (2020, July 28). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv.org. https://arxiv.org/abs/1910.10683

- Chat.openai.com. (n.d.). https://chat.openai.com/