Attention Redistribution During Event Segmentation In Large Language Model

Yuyang Xu
Department of Psychology
New York University
New York, NY 10012
yx2227@nyu.edu

Qingqing Yang
Department of Psychology
The Ohio State University
Columbus, OH 43210
yang.6118@osu.edu

Abstract

Human beings perceive a continuous string of experiences by segregating the experience into discrete events. Recently, it has been proven that a large language model can segregate events similarly to humans, even though the model is not specifically trained to do so. In this research, we used naturalistic stimuli like stories to explore the underlying changes in the attention mechanisms when large language model performs event segmentation. We discovered a redistribution of attention outputs toward words that play different roles in structuring an event. We found that the model enhances attention directed toward words indicative of potential changes in elements like time, space, objects, and goals in a continuous narrative. The model also reduces attention directed toward other kinds of words not indicative of such change. Our results provide better insights into the underlying processes of the high-level cognitive features in large language models and in the human brain.

1 Introduction

The stimulus received by the human sensory systems is continuous. Yet, when processing a long string of continuous incoming stimuli, humans often take the stimulus and perceive them as discrete events (Zacks et al., 2007). The ability to segment continuous stimuli into events is crucial to various cognitive processes (Jafarpour et al., 2022; Bangert et al., 2020; DuBrow and Davachi, 2016). Recently, it has been demonstrated that large language models (LLMs) like GPT-3 can segment events similarly to humans (Michelmann et al., 2023). Although the large language models are initially trained for next-word prediction, the ability of event segmentation emerges after extensive training naturally (Radford et al., 2019). This effect can be viewed as a convergent evolution; the challenge of language processing might force the human brains and LLMs to converge on similar properties (Waldrop, 2024). However, the exact mechanisms behind this capability's development remain an open question. Specifically, it is unclear how the changes in the inner function evolve during training.

In this work, we aim to provide additional insights into this question by looking into the attention mechanism of the LLM. We used the zero-shot prompting technique to compare the LLM attention activation pattern when it conducts event segmentation task versus not on various stories. The attention activation patterns are extracted at each layer and attention head (Vig and Belinkov, 2019). To examine how the model directs attention towards different event structures, we segregated the attention activities based on different grammatical features (nouns, verbs, determiners, etc.).

NeurIPS 2024 Workshop on Behavioral Machine Learning.

Table 1: Stories used

Story name	Number of events	Length (number of words)
Secret Life of Walter Mitty	6	1145
Story 1	8	817
Story 2	7	672
Story 3	8	673
Story 4	9	510
Story 5	9	663

2 Methods

2.1 Experimental procedures

An overview of our methods was shown in Figure 1 in supplementary material.

2.2 Text materials

We used six stories that involve continuous narration of physical and verbal interactions between the characters, as shown in Table 1. Our repository of stories consisted of one hand-written story and five GPT-generated stories. We used "Secret Life of Walter Mitty" for the hand-written story (Thurber, 1939). The natural structure of the story can be divided into separate, discrete events so that it was feasible for the model to perform the event segmentation task. We also used GPT-4 to generate five stories (Achiam et al., 2023). Each one of the stories included descriptions of sequential, real-life events that were structurally similar to the Walter Mitty story.

2.3 Prompts

To generate the story, we used the following prompt: "*Generate a story about a day in the life of college student Sarah. The story should be able to be divided into independent, discrete events. Shuffle the sequence of different events. Do not denote the event boundaries.*". When generating other stories, "*a day in the life of college student Sarah*" was replaced by other information and characters. Here, we specified that the story needs to be able to be divided into separate events instead of consisting of one continuous, inseparable event. In this way, it will be meaningful for the model to attempt event segmentation. Moreover, we asked GPT-4 to shuffle the events. By presenting the events out of their logical, linear sequence, the story ensured that the model did not rely solely on temporal cues to segment the events. Instead, it must rely on other indicators, such as changes in context, actions, or dialogue, to determine where one event ended and the other began.

To instruct the model to perform the event segmentation task, we leveraged the zero-shot prompting technique (Michelmann et al., 2023). The user role content prompt follows: "*An event is an ongoing coherent situation. The following story needs to be copied and segmented into events. Copy the following story word-for-word and start a new line whenever one event ends and another begins. This is the story:*". Then, the story will be inputted into the model. Additionally, we prompted to renew the instruction: "*This is a word-for-word copy of the same story that is segmented into events:*." The system role content was set as "*You are a person listening to a continuous story, which can be divided into distinct events.*".

In contrast, in the non-segmentation condition, we revised the prompt: "*Copy the following story word-for-word. This is the story:*". Then the story would also be inputted, and the model was prompted to renew the instruction: "*This is a word-for-word copy of the same story:*." The system role content was set as "*You are a person listening to a continuous story.*".

2.4 Model and grammatical feature extraction

The model we used to conduct event segmentation was `Llama-3-8B-Instruct` (Dubey et al., 2024). Stories were tokenized, and we used spaCy to extract grammatical features from the stories (Honnibal and Montani, 2017). Namely, we identified part-of-speech (POS) tags using spaCy. The POS tags T identified across all 6 stories were: T = {ADJ, ADP, ADV, AUX, DET, INTJ, NOUN, NUM, PART, PRON, PROPN, PUNCT, SCONJ, VERB}. Corresponding linguistic feature for each POS tag is shown in Table 2.

We extracted the attention output of each layer and each attention head, aggregated across each word token w_i in each POS tag. For a single POS tag T_j and layer k, the average attention output across all attention heads was given by:

$$\overline{A}_{j,k} = \frac{\sum_{w_i \in T_j} \sum_{h=1}^{H_k} S_{i,j,k,h}}{H_k} \tag{1}$$

where: $\overline{A}_{j,k}$ was the average attention output across attention heads for tag T_j in layer k. $S_{i,j,k,h}$ represents the attention output for word token w_i in tag T_j at layer k and head h. H_k was the number of attention heads in layer k. $S_{i,j,k,h}$ was formally defined as:

$$S_{i,j,k,h} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2}$$

where: Q was the query matrix. K was the key matrix. V was the value matrix. d_k was the dimensionality of the key vectors. softmax$(\cdot)$ was the softmax function applied row-wise (Vaswani, 2017).

We also calculated the attention output across all layers for each tag, which was formally defined as:

$$\text{Category Attention} = \sum_{k=1}^{L} \frac{\overline{A}_{j,k}}{L} \tag{3}$$

where: L was the total number of layers.

3 Results

3.1 Attention output across layers and attention heads

The attention towards per token identified in each POS tag at each layer and attention head was calculated with the formula for $S_{i,j,k,h}$. To compare the attention output across layers and attention heads across event segmentation conditions, a pairwise t-test was performed on flattened vectors containing 1024 measurements. The results showed that there were significant differences in attention scores for all the POS tags identified from stories ($p < 0.001$), as shown in Table 3 in the appendix. Therefore, the overall pattern of attention outputs for all the POS tags were sensitive to the event segmentation task. Most of the POS tags gained more attention in the event segmentation task, except that INTJ, NUM, and PROPN gained less attention attribution in event segmentation compared to the non-segmentation condition. The average attention outputs across all layers and attention heads for each POS tag per token were shown in Figure 30 in the appendix.

3.2 Attention pattern across layers

To assess whether the general layer-wise attention pattern was influenced by the event segmentation task, for each POS tag, we computed the pairwise t-test on attention scores among layers when averaging across attention heads. The t-test results were shown in Table 2. At the layer level, ADP, ADV, AUX, DET, PUNCT, SCONJ, and VERB gained significantly higher attention scores in the event segmentation task ($p < 0.05$). However, INTJ and PROPN gained lower attention scores in the event segmentation task ($p < 0.05$). There was no significant difference in the layer pattern of attention outputs between the two segmentation conditions for POS tag ADJ, NOUN, NUM, PART, and PRON ($p > 0.05$).

Additionally, to investigate whether the attention output of a specific layer was consistently sensitive to certain POS tags across stories, we performed a non-parametric Wilcoxon signed-rank test on each layer's averaged attention score between 2 segmentation conditions. None of the 32 layers exhibited significant differences for most of the POS tags. However, the model significantly and consistently attributed more attention to NOUN in layer 19 in the segmentation task compared to the non-segmentation condition ($W = 0$, $p = 0.0312$, $n = 6$). Besides, PUNCT gained significantly and consistently less attention in the event segmentation task in layers 2 and 11, while this category also

Table 2: T-test results for each POS across layers

POS Tag	Corresponding linguistic feature	t-value	df	p-value
ADJ	Adjective	1.99	31	0.0556
ADP	Adposition	5.78	31	0.0000
ADV	Adverb	6.32	31	0.0000
AUX	Auxiliary verb	4.89	31	0.0000
DET	Determiner	3.23	31	0.0029
INTJ	Interjection	−2.49	31	0.0181
NOUN	Noun	1.18	31	0.2488
NUM	Number	−2.00	31	0.0549
PART	Particle	1.54	31	0.1341
PRON	Pronoun	1.87	31	0.0712
PROPN	Proper noun	−2.65	31	0.0125
PUNCT	Punctuation	2.27	31	0.0306
SCONJ	Subordinating conjunction	4.98	31	0.0000
VERB	Verb	5.59	31	0.0000

consistently gained more attention in the event segmentation task across stories in layers 4, 14-15, 18-19, 21, 23-25 and 29 ($W = 0$, $p = 0.0312$, $n = 6$).

4 Discussion

In the sections above, we demonstrated that attention directed toward words like interjections, numbers, and proper nouns experienced reduction when LLM performs event segmentation. On the contrary, the attention directed toward other words like adverbs, verbs, and adpositions experienced a significant increase. Words like verbs, adverbs, and adpositions are essential in establishing relationships between words and constructing meanings in a sentence. For example, adpositions are crucial in indicating contextual information like the spatial and temporal relationships in narratives (Huddleston and Pullum, 2005). These categories of words are highly informative for discovering changes in elements like time, space, objects, and goals in a continuous narrative due to their semantic and grammatical roles in sentences (Ursini, 2011; Payne et al., 2010). Previous studies have shown that changes in these elements like time and space, are significantly correlated with the presence of event boundary detected by human (Michelmann et al., 2021). Therefore, it is possible that LLM redistributed more attention on words that indicate changes in the previously mentioned elements in the input text, contributing to the emerging feature of event segmentation. Although the initial training goal of LLM is not to perform event segmentation, the way it segments continuous events converges with the approach that human takes when doing the same task. In conclusion, our results provide an enhanced understanding of an naturally emerging cognitive feature, which is crucial in LLMs as well as in human high-level cognition. Moreover, previous research has demonstrated that LLMs can be considered a model organism for investigating language processing in the human brain (Tuckute et al., 2024; Goldstein et al., 2022; Kumar et al., 2022). Therefore, our results can potentially provide additional insights into underlying mechanism for continuous event processing in human, due to the naturally convergent property of the human brain and LLMs.

5 Limitations

Our task does not have relatively sufficient variations among story structures in terms of the total number of events and the total number of words. It would be ideal to examine our results further using stories that have drastically different total numbers of words and total number of events. Moreover, due to time and computing resources constraints, we only ran our task on one variation of the Llama 3 models. Testing with other language models across architectures and model sizes will be a promising research direction.

Acknowledgments and Disclosure of Funding

We thank Prof. Sebastian Michelmann at New York University and Prof. Hsin-Hung Li at The Ohio State University for their valuable inputs. We ran the model on New York University High Performance Computing Cluster (NYU HPC). We have not received any funding for this study. We declare no competing interests.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ashley S Bangert, Christopher A Kurby, Allyson S Hughes, and Omar Carrasco. Crossing event boundaries changes prospective perceptions of temporal length and proximity. *Attention, Perception, & Psychophysics*, 82:1459–1472, 2020.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Sarah DuBrow and Lila Davachi. Temporal binding within and across events. *Neurobiology of learning and memory*, 134:107–114, 2016.

Ariel Goldstein, Eric Ham, Samuel A Nastase, Zaid Zada, Avigail Grinstein-Dabus, Bobbi Aubrey, Mariano Schain, Harshvardhan Gazula, Amir Feder, Werner Doyle, et al. Correspondence between the layered structure of deep language models and temporal structure of natural language processing in the human brain. *BioRxiv*, pages 2022–07, 2022.

Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420, 2017.

Rodney Huddleston and Geoffrey Pullum. Introduction to english grammar. *Zeitschrift für Anglistik und Amerikanistik*, 53(2):195–197, 2005.

Anna Jafarpour, Elizabeth A Buffalo, Robert T Knight, and Anne GE Collins. Event segmentation reveals working memory forgetting rate. *Iscience*, 25(3), 2022.

Sreejan Kumar, Theodore R Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A Norman, Thomas L Griffiths, Robert D Hawkins, and Samuel A Nastase. Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. *BioRxiv*, pages 2022–06, 2022.

Sebastian Michelmann, Amy R Price, Bobbi Aubrey, Camilla K Strauss, Werner K Doyle, Daniel Friedman, Patricia C Dugan, Orrin Devinsky, Sasha Devore, Adeen Flinker, et al. Moment-by-moment tracking of naturalistic learning and its underlying hippocampo-cortical interactions. *Nature communications*, 12(1): 5394, 2021.

Sebastian Michelmann, Manoj Kumar, Kenneth A Norman, and Mariya Toneva. Large language models can segment narrative events similarly to humans. *arXiv preprint arXiv:2301.10297*, 2023.

John Payne, Rodney Huddleston, and Geoffrey K. Pullum. The distribution and category status of adjectives and adverbs. *Word Structure*, 3(1):31–81, 2010.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

James Thurber. "the secret life of walter mitty"(1939). 1939.

Greta Tuckute, Nancy Kanwisher, and Evelina Fedorenko. Language in brains, minds, and machines. *Annual Review of Neuroscience*, 47, 2024.

Francesco-Alessio Ursini. *The Language Of Space: The Acquisition And Interpretation of Spatial Adpositions In English*. PhD thesis, Macquarie University Printing services, 2011.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*, 2019.

MM Waldrop. Can chatgpt help researchers understand how the human brain handles language? *Proceedings of the National Academy of Sciences of the United States of America*, 121(25):e2410196121–e2410196121, 2024.

Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and Jeremy R Reynolds. Event perception: a mind-brain perspective. *Psychological bulletin*, 133(2):273, 2007.

A Appendix / supplemental material

A.1 The experimental process

The experimental process was shown in Figure 1.

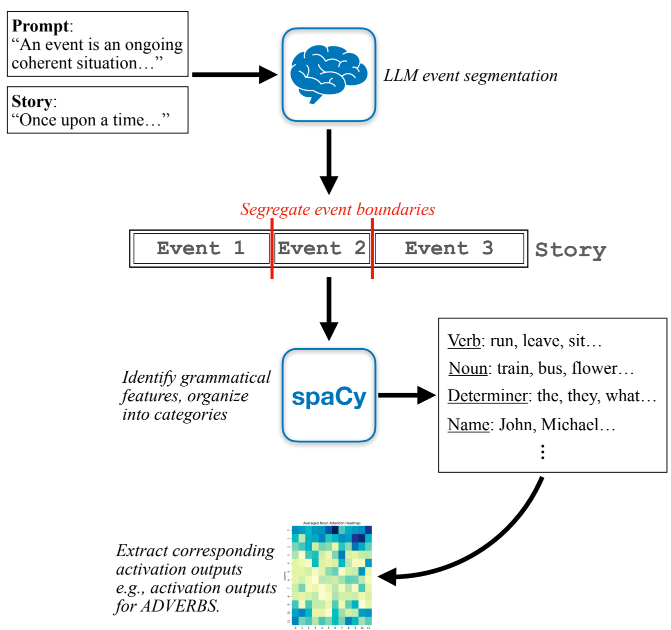

Figure 1: The experimental process

A.2 Code and datasets availability

All of the code and datasets required to reproduce our study are uploaded and can be accessed using the following link: `https://github.com/Qingqing-Yang-177/attention_event_llm/tree/main`

A.3 Attention distribution for each POS tag

The attention scores of each noun token (NOUN) across all the layers and attention heads were shown in Figure 2, broken out by layer (vertical axis) and head (horizontal axis). The attention scores of nouns at each layer were also shown in Figure 3. The attention was attributed more to nouns in middle layers, such as 9, 10, 14, and 15, and several attention heads are sensitive to nouns, such as layer 10 head 3, layer 14 head 4, layer 15 head 14, layer 16 head 27, and layer 17 head 9.

The same analysis and plots were performed and shown for all the POS tags. For adjective (ADJ) tokens, Figure 4 and Figure 5 indicated that the middle layers generally attributed more attention to them, such as layers 10, 12, 14, and 15. The layer 18 attention head 14 was specifically sensitive to adjective tokens.

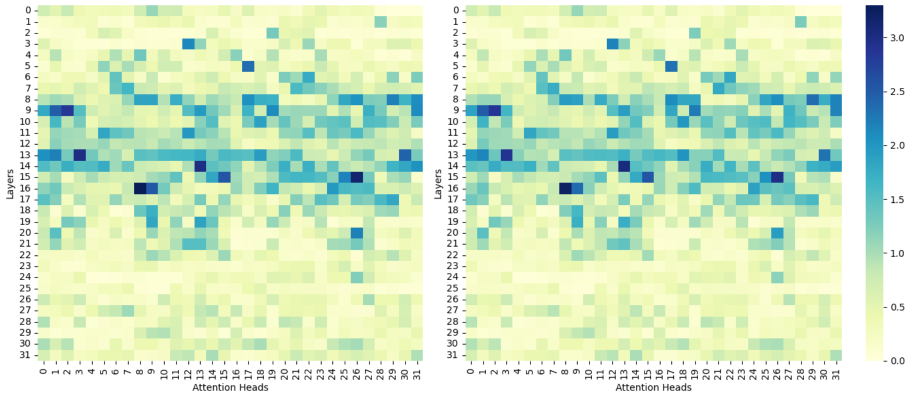

Figure 2: Heatmap of NOUN attention score in segmentation (left) and non-segmentation (right) task

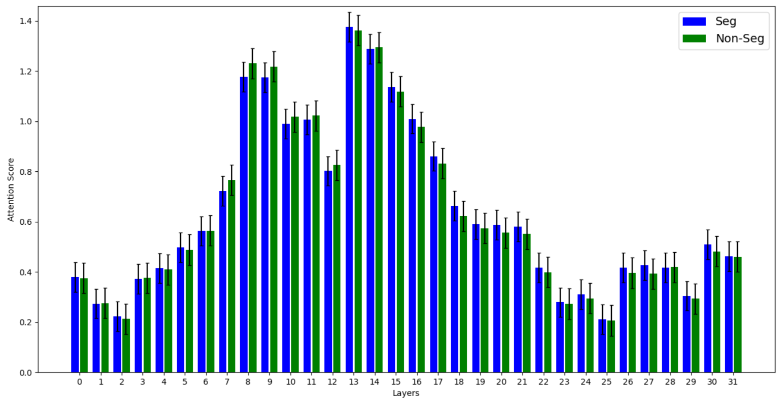

Figure 3: Layer average of NOUN attention score in two segmentation conditions

For adposition (ADP) tokens, Figure 6 and Figure 7 indicated that the middle layers generally attributed more attention to them, such as layers 12 and 14. Layer 13, head 2, and Layer 21, head 21 were sensitive to the adposition.

For adverb (ADV) tokens, Figure 8 and Figure 9 indicated that the middle layers generally attributed more attention to them, such as layers 14-15. Layer 13 head 2, layer 14 head 3, and Layer 18 head 14 were specifically sensitive to adverb tokens.

For auxiliary verb (AUX) tokens, Figure 10 and Figure 11 indicated that late middle layer 31 attention head 18 was specifically sensitive to auxiliary verb tokens.

For determiner (DET) tokens, Figure 12 and Figure 13 indicated that the early layer 1 and middle layers 12 and 14 generally attributed more attention to them. Layer 1, head 9, 11, and 16, and Layer 25, head 24 were specifically sensitive to determiner tokens.

For interjection (INTJ) tokens, Figure 14 and Figure 15 indicated that the early layer 1 head 10, 12, and 16, late layers 23 head 3 and layer 26 head 6 were specifically sensitive to interjection tokens.

For number (NUM) tokens, Figure 16 and Figure 17 indicated that the early layer 1 and middle layers 10, 14, 15, and 16 generally attributed more attention to them. The layer 1 head 11, layer 10 head 28, layer 14 head 2, 13, and 31, layer 21 head 26, layer 24 head 15, layer 29 head 29, and layer 30 head 17 were specifically sensitive to number tokens.

For particle (PART) tokens, Figure 18 and Figure 19 indicated that layer 1 generally attributed more attention to them, and late middle layer 31 attention head 18 were specifically sensitive to particle tokens, similar to Auxiliary verb (AUX).

For pronoun (PRON) tokens, Figure 20 and Figure 21 indicated that the early layer 1 and middle layers 14 and 15 generally attributed more attention to them. Layer 1, heads 9 and 16; layer 14, head 3; layer 17, head 2; layer 21, heads 16 and 24; layer 25, head 24; and Layer 31, head 18, were specifically sensitive to pronoun tokens.

For proper noun (PROPN) tokens, Figure 22 and Figure 23 indicated that the middle layers generally attributed more attention to them, such as layers 9, 10, 12, 14, and 15. The layer 17 attention head 14 was specifically sensitive to proper noun tokens.

For punctuation (PUNCT) tokens, Figure 24 and Figure 25 indicated that layer 1 generally attributed more attention to them. The layer 1 attention heads 29, 30, and 31 were specifically sensitive to punctuation tokens, similar to particle (PART) and auxiliary verbs (AUX).

For subordinating conjunction (SCONJ) tokens, Figure 26 and Figure 27 indicated that the middle layers generally attributed more attention to them, such as layers 9-12 and 14-16. The layer 13 attention heads 1 and 2, layer 21 head 21, and layer 29 head 26 were specifically sensitive to subordinating conjunction tokens.

For verb (VERB) tokens, Figure 28 and Figure 29 indicated that the middle layers generally attributed more attention to them, such as layers 9-12 and 14-16. The layer 1 head 11, layer 11 head 15, layer 14 head 14, layer 20 head 14, and layer 21 head 2 were specifically sensitive to verb tokens.

A.4 Attention score across layers and attention heads

Here we exhibited the pairwise t-test results in Table 3 for each POS tag when comparing the attention score pattern across layer and attention heads in two segmentation conditions.

Table 3: T-test for each POS across layers and attention heads

POS Tag	Corresponding linguistic feature	t-value	df	p-value
ADJ	Adjective	5.22	1023	0.0000
ADP	Adposition	10.16	1023	0.0000
ADV	Adverb	16.01	1023	0.0000
AUX	Auxiliary verb	7.64	1023	0.0000
DET	Determiner	6.81	1023	0.0000
INTJ	Interjection	-4.65	1023	0.0000
NOUN	Noun	3.56	1023	0.0000
NUM	Number	-4.68	1023	0.0000
PART	Particle	3.32	1023	0.0000
PRON	Pronoun	3.94	1023	0.0000
PROPN	Proper noun	-8.17	1023	0.0000
PUNCT	Punctuation	5.75	1023	0.0000
SCONJ	Subordinating conjunction	8.65	1023	0.0000
VERB	Verb	11.99	1023	0.0000

A.5 Attention score across all POS tags

The averaged attention score for each POS tag, across stories, layers, and attention heads, in two segmentation conditions is shown in Figure 30. The pairwise t-test was conducted to compare the overall attention scores across all POS tags between two segmentation conditions and investigate whether event segmentation increased or decreased the overall attention attributed to per token. The results indicate that the difference between the two conditions was not statistically significant ($t(13) = 1.96, p = 0.0721$).

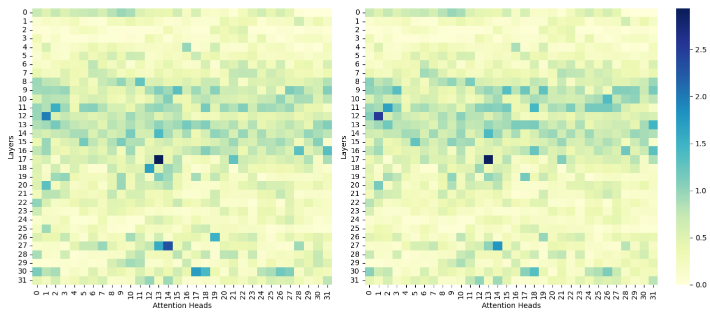

Figure 4: Heatmap of ADJ attention score in segmentation (left) and non-segmentation (right) task

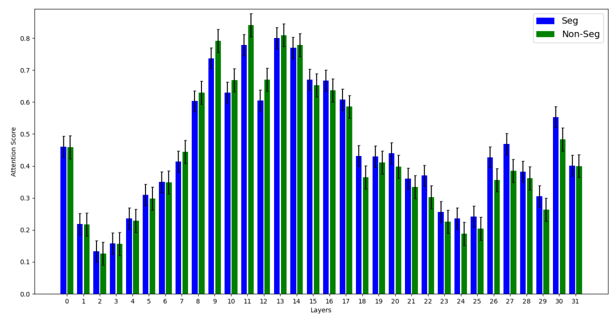

Figure 5: Layer average of ADJ attention score in two segmentation conditions

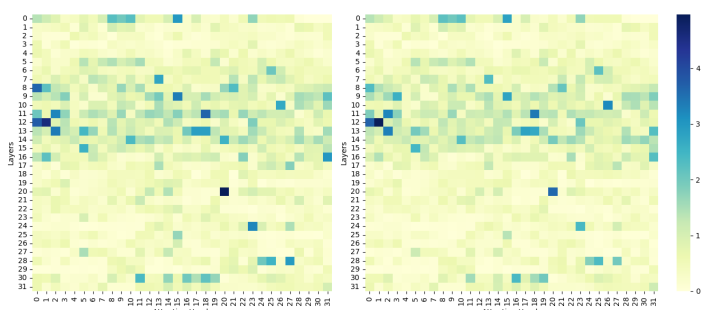

Figure 6: Heatmap of ADP attention score in segmentation (left) and non-segmentation (right) task

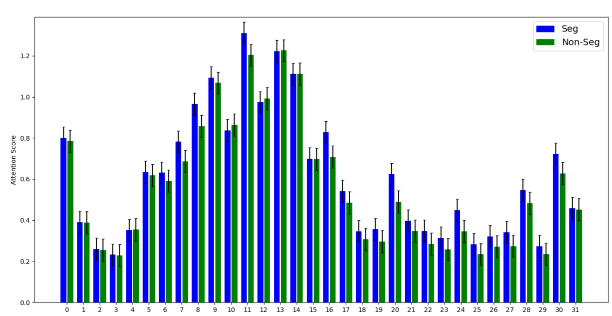

Figure 7: Layer average of ADP attention score in two segmentation conditions

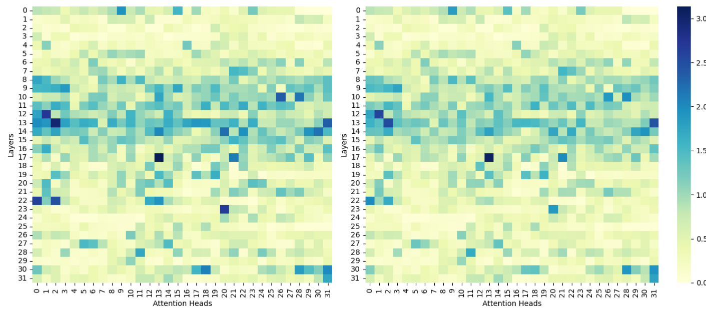

Figure 8: Heatmap of ADV attention score in segmentation (left) and non-segmentation (right) task

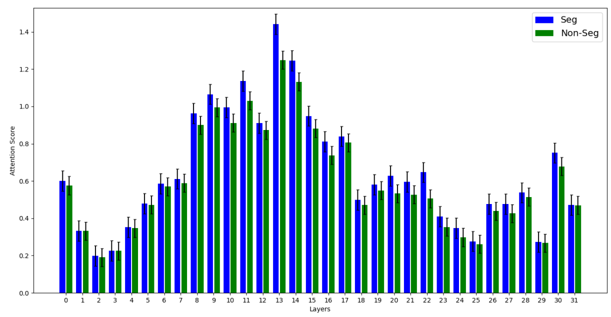

Figure 9: Layer average of ADV attention score in two segmentation conditions

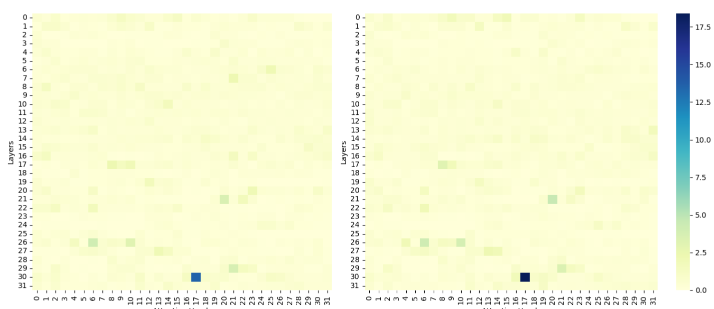

Figure 10: Heatmap of AUX attention score in segmentation (left) and non-segmentation (right) task

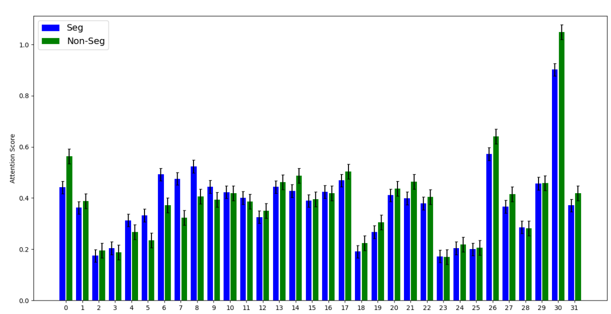

Figure 11: Layer average of AUX attention score in two segmentation conditions

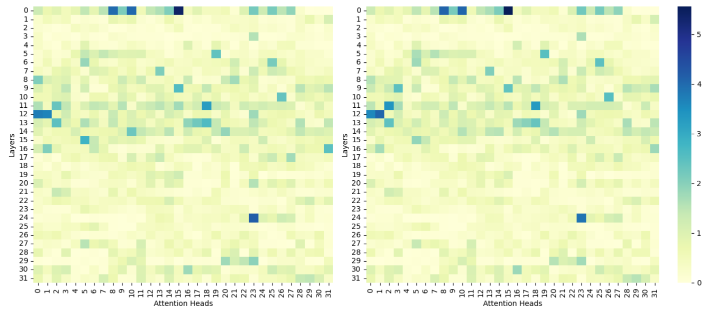

Figure 12: Heatmap of DET attention score in segmentation (left) and non-segmentation (right) task

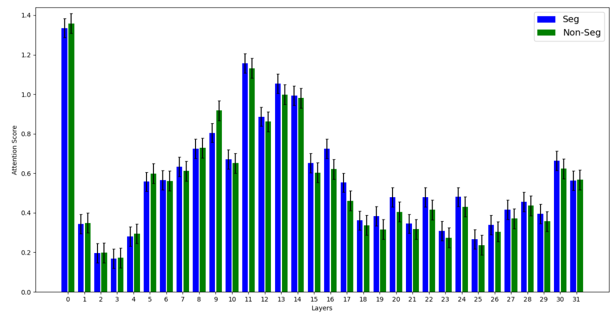

Figure 13: Layer average of DET attention score in two segmentation conditions

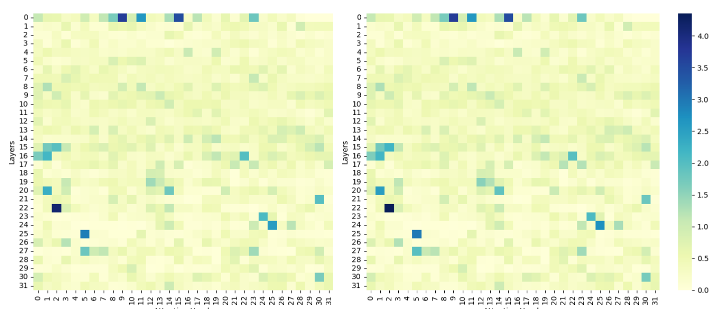

Figure 14: Heatmap of INTJ attention score in segmentation (left) and non-segmentation (right) task

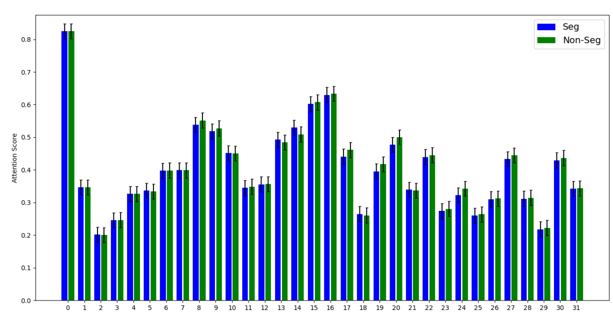

Figure 15: Layer average of INTJ attention score in two segmentation conditions

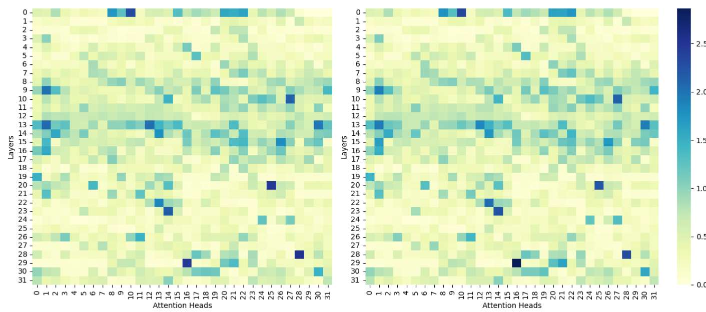

Figure 16: Heatmap of NUM attention score in segmentation (left) and non-segmentation (right) task

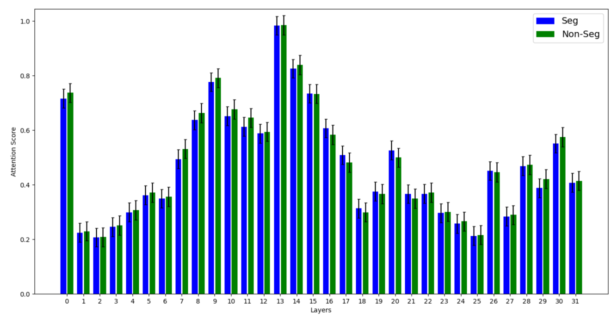

Figure 17: Layer average of NUM attention score in two segmentation conditions

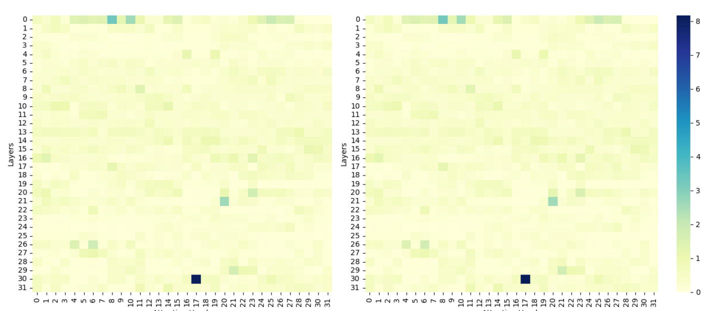

Figure 18: Heatmap of PART attention score in segmentation (left) and non-segmentation (right) task

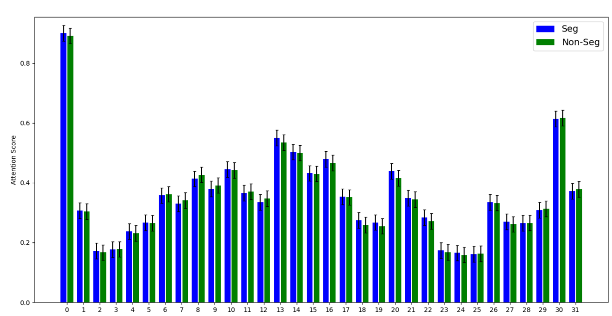

Figure 19: Layer average of PART attention score in two segmentation conditions

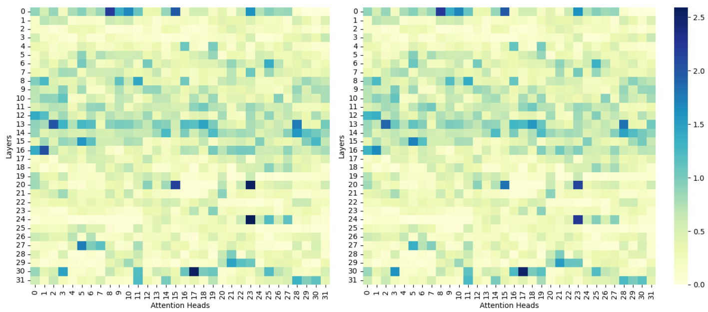

Figure 20: Heatmap of PRON attention score in segmentation (left) and non-segmentation (right) task

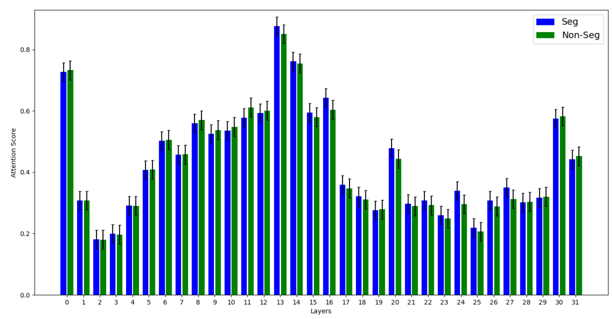

Figure 21: Layer average of PRON attention score in two segmentation conditions

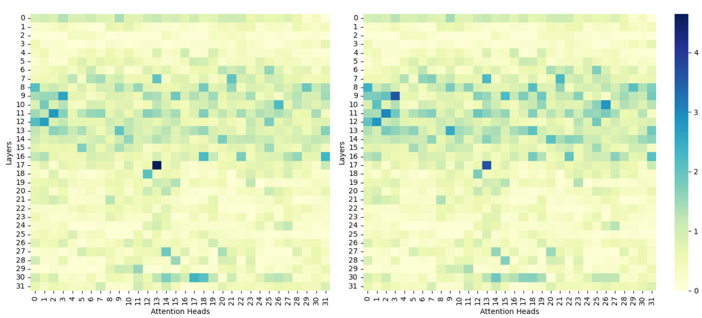

Figure 22: Heatmap of PROPN attention score in segmentation (left) and non-segmentation (right) task

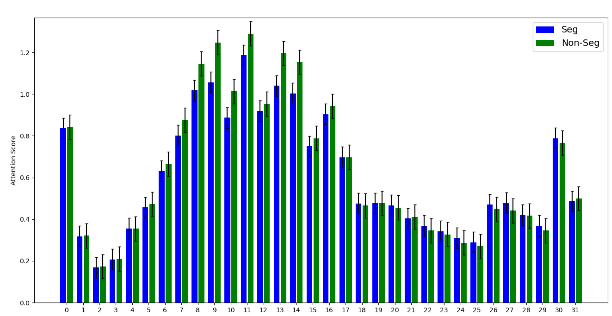

Figure 23: Layer average of PROPN attention score in two segmentation conditions

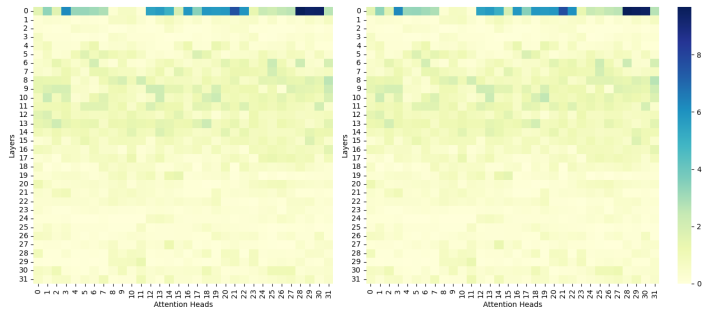

Figure 24: Heatmap of PUNCT attention score in segmentation (left) and non-segmentation (right) task

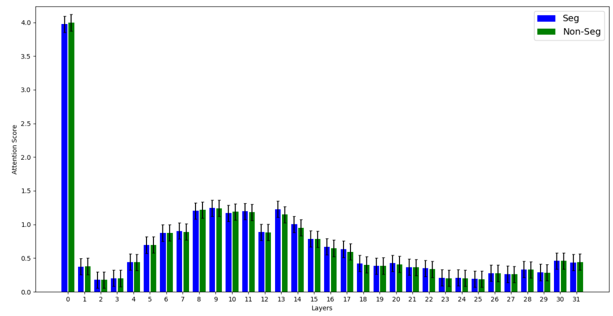

Figure 25: Layer average of PUNCT attention score in two segmentation conditions

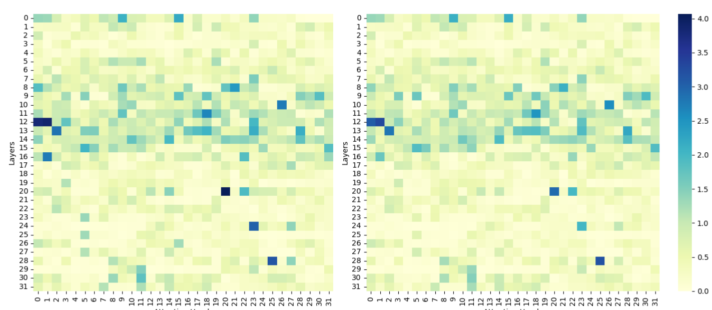

Figure 26: Heatmap of SCONJ attention score in segmentation (left) and non-segmentation (right) task

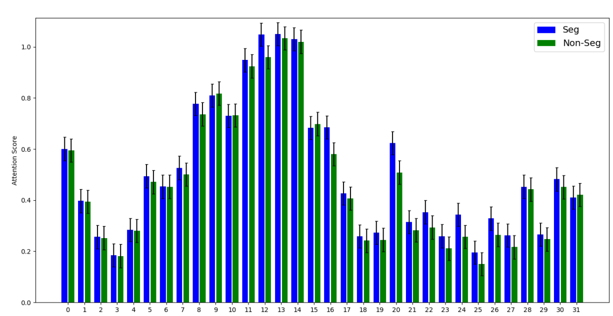

Figure 27: Layer average of SCONJ attention score in two segmentation conditions

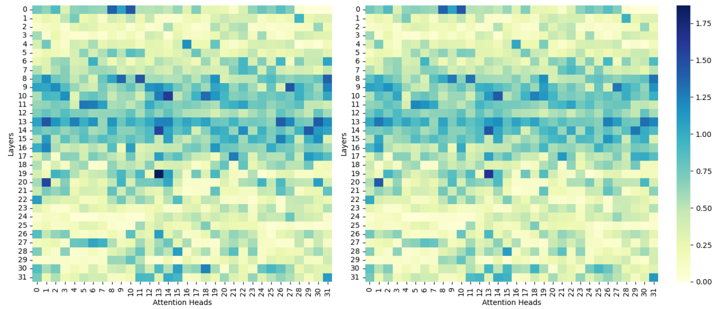

Figure 28: Heatmap of VERB attention score in segmentation (left) and non-segmentation (right) task

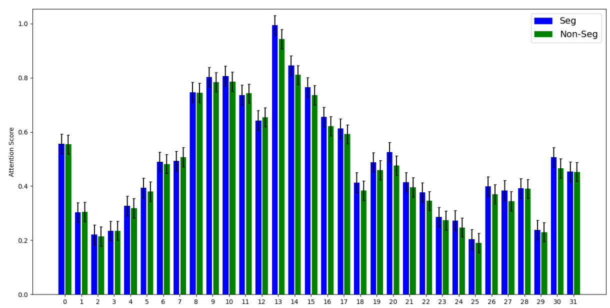

Figure 29: Layer average of VERB attention score in two segmentation conditions

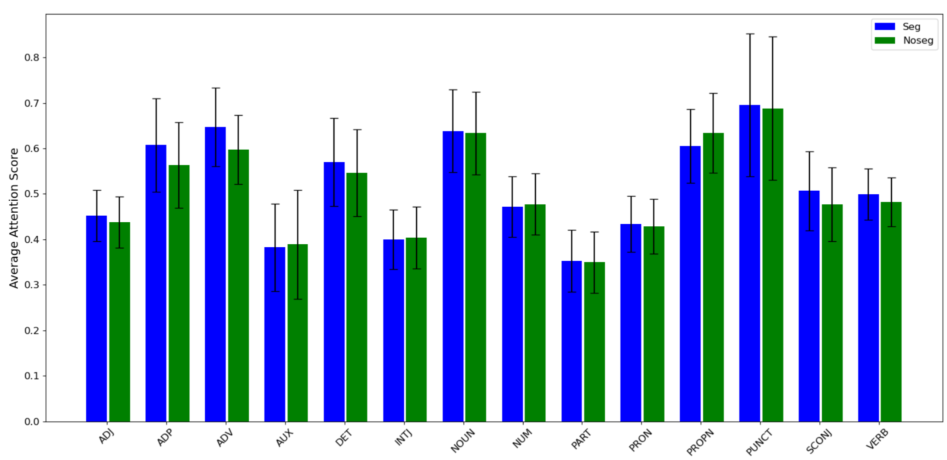

Figure 30: POS tag attention.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please refer to the abstract, introduction, and discussion sections of this paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to the limitation section of this paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Please refer to the method section.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please refer to the methods section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Please refer to the GitHub repository link in the manuscript. The repository contains the scripts (code) and the story (data).

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to the methods section, and also the scripts folder in the GitHub repository.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Please refer to the results section and appendix. The calculation processes is also provided in the scripts folder of the GitHub repository.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to the methods section and the bash file in the scripts folder in the GitHub repository.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to the discussion section of the paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Please refer to the reference section of the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.